New England Common Assessment Program

2005–2006 Technical Report

September 2006

Measured Progress

100 Education Way, Dover, NH 03820 (800) 431-8901 www.measuredprogress.org

TABLE OF CONTENTS

CHAPTER 1 – OVERVIEW

1.1 PURPOSE OF THE NEW ENGLAND COMMON ASSESSMENT PROGRAM

The New England Common Assessment Program (NECAP) is the result of collaboration among New Hampshire (NH), Rhode Island (RI), and Vermont (VT) to build a set of assessments for grades 3 through 8 to meet the requirements of the No Child Left Behind Act (NCLB). The purposes of the tests are as follows: (1) Provide data on student achievement in reading/language arts and mathematics to meet the requirements of NCLB; (2) provide information to support program evaluation and improvement; and (3) provide to parents and the public information on the performance of students and schools. Therefore, the tests are constructed to meet rigorous technical criteria, include universal design elements and accommodations so that students can access test content, and gather reliable student demographic information for accurate reporting. School improvement is supported by

- providing a transparent test design through the grade-level expectations (GLEs), distributions of emphasis, and practice tests;
- reporting results by GLE subtopics, released items, and subgroups; and
- hosting test interpretation workshops to foster understanding of results.

Student-level results are provided to schools and families to be used as one piece of evidence about progress and learning that occurred on the prior year's GLEs. The results are a status report of a student's performance against GLEs and should be used cautiously in partnership with local data.

1.2 PURPOSE OF THIS REPORT

The purpose of this technical report is to document the technical aspects of the 2005–2006 NECAP. In October of 2005, students in grades 3 through 8 participated in the administration of the NECAP in reading and mathematics. Students in grades 5 and 8 also participated in writing. This report provides information about the technical quality of those assessments, including a description of the processes used to develop, administer, and score the tests and to analyze the test results. This report is intended to serve as a guide for replicating and/or improving the procedures in subsequent years.

Though some parts of this technical report may be used by educated laypersons, the intended audience is experts in psychometrics and educational research. The report assumes a working knowledge of measurement concepts, such as "reliability" and "validity," and statistical concepts, such as "correlation" and "central tendency." In some chapters, the reader is presumed also to have basic familiarity with advanced topics in measurement and statistics.

1.3 ORGANIZATION OF THIS REPORT

The organization of this report is based on the conceptual flow of an assessment's life span; the report begins with the initial test specification and addresses all the intermediate steps that lead to final score reporting. Section I provides a description of the NECAP test. It consists of four chapters covering the test design and development process, the administration of the tests, scoring, and equating and scaling. Section II provides the Statistical and Psychometric Summaries. It consists of three chapters covering item analysis, reliability, and validity. Section III covers the 2005 NECAP score reporting. Section IV contains references, and Section V contains the appendices.

SECTION I—DESCRIPTION OF THE 2005 NECAP TEST

CHAPTER 2—DEVELOPMENT AND TEST DESIGN

2.1 DEVELOPMENT OF 2004 PILOT TESTS

In preparation for the first operational administration of the NECAP in October of 2005, a pilot test was conducted in 2004, with the following purposes:

- Field-test all newly developed reading and mathematics items to be used in the common and matrix-equating sections of the following year's operational test.
- Field-test, construct, and pre-equate the operational writing forms for all years of the program; this was done to avoid the need for ongoing field-testing of extended writing tasks in the embedded field-test section of the operational test.
- Try out all procedures and materials of the program (e.g., the timing of test sessions; accommodations; test administrator and test coordinator manuals; mathematics tool kits; shipping/receiving processes; and the like) before the first operational administration.
- Provide to schools the opportunity to experience the new assessment so as to assist them in getting prepared for the first operational administration.
- Obtain feedback from students, test administrators, and test coordinators in order to make any necessary modifications.

The test development process for the pilot test mirrored the process described in this chapter for the operational test. The number of items developed and field-tested are listed on the following page.

Reading: Grades 3–8

	Populate first year's forms (not counting embedded FT)	Initial FT	Items Developed
Passages	5 long 5 short	8 long 8 short	10 long 10 short
MC	40 long 20 short	80 long 48 short	100 long 60 short
CR	10 long 5 short	24 long 16 short	30 long 20 short
Stand-Alone MC	10	16	20

Mathematics: Grades 3 & 4

	Populate first year's forms (not counting embedded FT)	Initial FT	Items Developed
MC	71	105	140
SA1	22	30	40
SA2	22	30	40

Mathematics: Grades 5–8

	Populate first year's forms (not counting embedded FT)	Initial FT	Items Developed
MC	68	108	134
SA1	18	27	36
SA2	18	27	36
CR	10	12	20

Writing: Grades 5 & 8

	Needed for 4 operational forms	FT ~150%	Develop ~200%
MC	60	90	120
CR	12	18	24
ER	4	8	12

TEST DESIGN OF THE 2004 PILOT TESTS

Because one of the purposes of the pilot test administration was to give schools an opportunity to experience what the operational test would be like, the pilot test forms were constructed to mirror the intended test design. The only difference was that all item positions on the pilot forms were populated with field-test items. The designs of the pilot tests are presented on the following pages.

Reading: Grades 3–8:

- 6 forms, 3 blocks (and sessions) each
- Block = 1 long, 1 short, and 2 stand-alone MC (14 MC, 3 CR)
- Each passage is repeated in two forms: 10 unique MC and 3 unique CR for each long passage and 6 unique MC and 2 unique CR for each short passage

	Blocks															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Form 1	X	X	X													
Form 2				X	X	X										
Form 3							X	X	X							
Form 4										X	X	X				
Form 5													X	X	X	
Form 6	X	X														X

Note: Word ID MC needs 8 blocks (repeat in blocks 9–16).

Block Descriptions

Block 1 Long Passage A – MC #1–8 and CR#1–2
Short Passage A – MC#1–4 and CR#1
2 Word ID MC

Block 2 Long Passage B – MC#1–8 and CR#1–2
Short Passage B – MC#1–4 and CR#1
2 Word ID MC

Block 3 Long Passage C – MC#1–8 and CR#1–2
Short Passage C – MC#1–8 and CR#1
2 Word ID MC

Block 4 Long Passage D – MC#1–8 and CR#1–2
Short Passage D – MC#1–4 and CR#1
2 Word ID MC

Block 5 Long Passage E – MC#1–8 and CR#1–2
Short Passage E – MC#1–4 and CR#1
2 Word ID MC

Block 6 Long Passage F – MC#1–8 and CR#1–2
Short Passage F – MC#1–4 and CR#1
2 Word ID MC

Block 7 Long Passage G – MC#1–8 and CR#1–2
Short Passage G – MC#1–4 and CR#1
2 Word ID MC

Block 8 Long Passage H – MC#1–8 and CR#1–2
Short Passage H – MC#1–4 and CR#1
2 Word ID MC

Block 9 Long Passage A – MC#3–10 and CR#2–3
Short Passage A – MC#3–6 and CR#2
2 Word ID MC (repeated)

Block 10 Long Passage B – MC#3–10 and CR#2–3
Short Passage B – MC#3–6 and CR#2
2 Word ID MC (repeated)
Block 11 Long Passage C – MC#3–10 and CR#2–3
Short Passage C – MC#3–6 and CR#2
2 Word ID MC (repeated)
Block 12 Long Passage D – MC#3–10 and CR#2–3
Short Passage D – MC#3–6 and CR#2
2 Word ID MC (repeated)
Block 13 Long Passage E – MC#3–10 and CR#2–3
Short Passage E – MC#3–6 and CR#2
2 Word ID MC (repeated)
Block 14 Long Passage F – MC#3–10 and CR#2–3
Short Passage F – MC#3–6 and CR#2
2 Word ID MC (repeated)
Block 15 Long Passage G – MC#3–10 and CR#2–3
Short Passage G – MC#3–6 and CR#2
2 Word ID MC (repeated)
Block 16 Long Passage H – MC#3–10 and CR#2–3
Short Passage H – MC#3–6 and CR#2
2 Word ID MC (repeated)

Mathematics: Grades 3 & 4

- 3 forms, 3 blocks (and sessions) each
- Block = 15 MC, 4 SA-1, 4 SA-2

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8
Form 1	X	X	X					
Form 2				X	X	X		
Form 3	X						X	X

Notes: MC needs 7 blocks (repeat 1).
SA1 and SA2 need 8 blocks with 2 items repeated.

Mathematics: Grades 5–8

- 3 forms, 3 blocks (and sessions) each
- Block = 12 MC, 3 SA1, 3 SA2, 2 CR

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7	Block 8	Block 9
Form 1	X	X	X						
Form 2				X	X	X			
Form 3							X	X	X

Notes: MC needs 9 blocks.
SA1 and SA2 need 9 blocks (repeat 1).
CR needs 6 blocks (repeat 3).

Writing: Grades 5 & 8
-16 overlapping forms
-6 Unique A Blocks = 10MC and 3 CR
-8 Unique B Blocks = 5 MC and 1 ER

Form	A Block	B Block
1	A1	B1
2	A1	B2
3	A2	B2
4	A2	B3
5	A3	B3
6	A3	B4
7	A4	B4
8	A4	B5
9	A5	B5
10	A5	B6
11	A6	B6
12	A6	B7
13	A1	B7
14	A1	B8
15	A2	B8
16	A2	B1

SAMPLING PLAN FOR THE 2004 PILOT TESTS

All schools and all students in grades 3 through 8 participated in the pilot test (one exception was that Braille forms were not produced for the pilot test). A complete test (three sessions) of a single content area was administered to each student. A form of each content area was assigned to schools by grade level so as to ensure that all schools would experience all content areas across the grades in the school, but each school was assigned only one from of a content area to minimize exposure of the forms in any one school.

Forms were assigned to schools to avoid any difference between forms with respect to schools' previous performance on state assessments in each of the states and across the three states. Each test form was administered to approximately 1200.

Preliminary form assignment was accomplished by the following procedures for each grade within each state:

1. Determine the maximum number of students that can be assigned to each form (Form Maximum). This is the number of students in the grade divided by the number of forms to assign, plus the standard deviation of the school enrollment.
2. Sort the data by descending enrollment.
3. Randomly determine the starting form number
4. Make the following comparison before assigning the form to a record:

   Is the grade enrollment plus number of students already assigned to the form greater than the Form Maximum?

   a. If no, assign for the record.

   b. If yes, increase the form number by one and perform the same calculation
5. Increase the form number by 1 and move to the next record

   After all grades are assigned a form for each state, an analysis/reassignment is done to ensure that each school with two or more grades is taking at least two subjects.

1. All schools with two or more grades taking only one subject are identified.
2. One grade is chosen from each school to have the subject (form) reassigned. To identify which grade within a school is reassigned, the following statistic is calculated:

   Rank Variable = (Form Max–number of students assigned to form) / grade enrollment for school

   The grade with the smallest Rank Variable value is selected for reassignment.
3. The reassigned records are processed by state and by grade. The records are sorted again in descending order of enrollment.
4. To determine how the forms are reassigned, the available forms are sorted in ascending order by the number of students assigned to the form.

5. The first record is assigned to the smallest form, the second record to the second smallest form, and so forth. The form order is repeated as many times as is necessary depending on the number of records within the grade needing reassignment.

Grades 5 and 8 (due to the additional testing of writing) were excluded from the reassignment process because the number of students assigned to each form during the preliminary assignment was already near the threshold number of students needed at each form for scoring.

SCORING OF THE 2004 PILOT TESTS

All student responses to MC questions were scanned and included in the analysis of the items (sometimes in excess of the 1200 as stated). All available SA, CR, and ER items were benchmarked and scored up to the threshold of 1200.

Because the pilot test was conducted to emulate the subsequent operational test as much as possible, readers are referred to other chapters of this report for more specific details.

2.2 OPERATIONAL DEVELOPMENT PROCESS

GRADE-LEVEL EXPECTATIONS

NECAP test items are directly linked to the **content standards** and **performance indicators** described in the GLEs. The content standards for each grade are grouped into content cluster levels for purposes of reporting results; the performance indicators are used by content specialists to help to guide the development of test questions. An item may address one, several, or all of the performance indicators.

EXTERNAL ITEM REVIEW

IRCs were formed by the states to provide an external review of the items. The committees are made up of teachers, curriculum supervisors, and higher-education faculty from the states, and all committee members serve rotating terms. A list of IRC member names and affiliations is included in

Appendix A. The committees' primary role is to review test items for the NECAP, provide feedback on the items, and make recommendations on which items should be selected for use in the program. The 2005–06 NECAP IRCs for each content area in grade levels 3 through 8 met twice. In the first meeting, committee members reviewed proposed reading passages and samples of test items and scoring rubrics for the embedded field test that would fill the gaps left in coverage of the standards after items moved to common. In the second meeting, committee members reviewed the entire set of the embedded field-test items proposed for the 2005–06 operational test and made recommendations about selecting, revising, or eliminating specific items from the item pool for the operational test. During the meeting, the members were asked to review each item against the following criteria:

- **Grade-Level Expectation Alignment**

  - Is the test item aligned to the appropriate GLE?

  - If not, which GLE or grade level is more appropriate?

- **Correctness**

  - Are the items and distracters correct with respect to content accuracy and developmental appropriateness?

  - Are the scoring guides consistent with GLE wording and developmental appropriateness?

- **Depth of Knowledge***

  - Are the items coded to the appropriate Depth of Knowledge?

  - If consensus cannot be reached, is there clarity around why the item might be on the borderline of two levels?

  * NECAP employed the work of Dr. Norman Webb to guide the development process with respect to Depth of Knowledge. Dr. Webb conducted a training workshop for staff from Measured Progress and the state departments of education. Test specification documents identified ceilings and targets for Depth of Knowledge codings.

- **Language**

  - Is the item language clear?

  - Is the item language accurate (syntax, grammar, conventions)?

- **Universal Design**

  - Is there an appropriate use of simplified language (does not interfere with the construct being assessed)?

  - Are charts, tables, and diagrams easy to read and understandable?

  - Are charts, tables, and diagrams necessary to the item?

  - Are instructions easy to follow?

  - Is the item amenable to accommodations—read aloud, signed, or Braille?

INTERNAL ITEM REVIEW

- The lead Measured Progress test developer within the content specialty reviewed the formatted item, CR scoring guide, and any reading selections and graphics.
- The content reviewer considered item "integrity," item content and structure, appropriateness to designated content area, item format, clarity, possible ambiguity, answer cueing, appropriateness and quality of reading selections and graphics, and appropriateness of scoring guide descriptions and distinctions (as correlated to the item and within the guide itself). The item reviewer also ensured that, for each item, there was only one correct answer.
- The content reviewer also considered scorability and evaluated whether the scoring guide adequately addressed performance on the item.
- Fundamental questions that the content reviewer considered, but was not limited to, included the following:

  - What is the item asking?

  - Is the key the only possible key? (Is there only *one* correct answer?)

- Is the CR item scorable as written (were the correct words used to elicit the response defined by the guide)?
- Is the wording of the scoring guide appropriate and parallel to the item wording?
- Is the item complete (e.g., with scoring guide, content codes, key, grade level, and contract identified)?
- Is the item appropriate for the designated grade level?

BIAS AND SENSITIVITY REVIEW

Bias review is an essential component of the development process. During the bias review process, NECAP items were reviewed by a committee of teachers, English language learner (ELL) specialists, special-education teachers, and other educators and members of major constituency groups who represent the interests of legally protected and/or educationally disadvantaged groups. A list of bias and sensitivity review committee member names and affiliations are included in Appendix A. Items were examined for issues that might offend or dismay students, teachers, or parents. Including such groups in the development of assessment items and materials can avoid many unduly controversial issues, and unfounded concerns can be allayed before the test forms are produced.

ITEM EDITING

Measured Progress editors reviewed and edited the items to ensure uniform style (based on *The Chicago Manual of Style,* 14th edition) and adherence to sound testing principles. These principles included the stipulation that items

- were correct with regard to grammar, punctuation, usage, and spelling;
- were written in a clear, concise style;
- contained unambiguous explanations to students as to what is required to attain a maximum score;
- were written at a reading level that would allow the student to demonstrate his or her knowledge of the tested subject matter, regardless of reading ability;

- exhibited high technical quality regarding psychometric characteristics;
- had appropriate answer options or score-point descriptors; and
- were free of potentially sensitive content.

REVIEWING AND REFINING

Test developers presented item sets to the development committees for their recommendations for placement of items into the embedded field-test portions of the test. The NH, RI, and VT Departments of Education content specialists made the final selections with the assistance of Measured Progress at a final face-to-face meeting.

OPERATIONAL TEST ASSEMBLY

At Measured Progress, test assembly is the sorting and laying out of item sets into test forms. Criteria considered during this process included the following:

- **Content coverage/match to test design.** The Measured Progress curriculum and assessment specialists completed an initial sorting of items into sets based on a balance of content categories across sessions and forms, as well as a match to the test design (e.g., number of MC, SA, and CR items).
- **Item difficulty and complexity.** Item statistics drawn from the data analysis of previously tested items were used to ensure similar levels of difficulty and complexity across forms.
- **Visual balance.** Item sets were reviewed to ensure that each reflected a similar length and "density" of selected items (e.g., length/complexity of reading selections, or number of graphics).
- **Option balance.** Each item set was checked to verify that it contained a roughly equivalent number of key options (A, B, C, and D).
- **Name balance.** Item sets were reviewed to ensure that a diversity of student names was used.
- **Bias.** Each item set was reviewed to ensure fairness and balance based on gender, ethnicity, religion, socioeconomic status, and other factors.

- **Page fit.** Item placement was modified to ensure the best fit and arrangement of items on any given page.
- **Facing-page issues.** For multiple items associated with a single stimulus (a graphic or reading selection), consideration was given both to whether those items needed to begin on a left- or right-hand page to the nature and amount of material that needed to be placed on facing pages. These considerations served to minimize the amount of "page flipping" required of students.
- **Relationship between forms.** Although embedded field-test items differ from form to form, they must take up the same number of pages in each form so that sessions and content areas begin on the same page in every form. Therefore, the number of pages needed for the longest form often determines the layout of each form.
- **Visual appeal.** The visual accessibility of each page of the form was always taken into consideration, including such aspects as the amount of "white space," the density of the text, and the number of graphics.

EDITING DRAFTS OF OPERATIONAL TESTS

Any changes made by a test construction specialist must be reviewed and approved by a test developer. After a form had been laid out in what was considered its final form, it was reread to identify any final considerations, including the following:

- Editorial changes. All text was scrutinized for editorial accuracy, including consistency of instructional language, grammar, spelling, punctuation, and layout. Measured Progress's publishing standards are based on *The Chicago Manual of Style,* 14th edition.
- "Keying" items. Items were reviewed for any information that might "key" or provide information that would help to answer another item. Decisions about moving keying items are based on the severity of the "key-in" and the placement of the items in relation to each other within the form.

- Key patterns. The final sequence of keys was reviewed to ensure that their order appeared random (e.g., no recognizable pattern and no more than three of the same key in a row).

BRAILLE AND LARGE-PRINT TRANSLATION

Common items for grades 3 through 8 were translated into Braille by a subcontractor that specializes in test materials for blind and visually impaired students. In addition, Form 1 for each grade was also adapted into a large-print version.

2.3 ITEM TYPES

NH, RI, and VT educators and students were familiar with the item types that were used in the 2005–06 assessment, as all had been previously introduced on the pilot tests administered in October of 2004. The item types used and the functions of each are described below.

Multiple-Choice (MC) items were administered in grades 3 through 8 in reading, mathematics, and writing to provide breadth of coverage of the GLEs. Because they require approximately one minute for most students to answer, these items make efficient use of limited testing time and allow coverage of a wide range of knowledge and skills.

Short-Answer (SA) items were administered in grades 3 through 8 mathematics only to assess students' skills and their abilities to work with brief, well-structured problems that had one solution or a very limited number of solutions. SA items require approximately two to five minutes for most students to answer. The advantage of this item type is that it requires students to demonstrate knowledge and skills by generating, rather than merely selecting, an answer.

Constructed-Response (CR) items typically require students to use higher-order thinking skills—evaluation, analysis, summarization, and so on—in constructing a satisfactory response. CR items should take most students approximately five to ten minutes to complete. These items were administered in grades 3 through 8 in reading, in grades 5 and 8 in writing, and in grades 5 through 8 in mathematics

A single common **writing prompt** with three SA planning box items was administered in grades 5 and 8. Students were given 45 minutes (plus limited additional time if necessary) to compose a response that was scored by two independent readers both on the quality of the stylistic and rhetorical aspects of the writing and on as the use of standard English conventions. Students were encouraged to write a rough draft and were advised by the test administrator when to begin copying their final draft into their student answer booklets.

Approximately twenty-five percent of the common NECAP items were released to the public in 2005–06, and the goal is to increase that percentage to fifty percent in 2006–07 and beyond. The released NECAP items are posted on a Web site hosted by Measured Progress and on the Department of Education Web sites. Schools are encouraged to incorporate the use of released items in their instructional activities so that students will be familiar with them.

2.4 Operational Test Designs and Blueprints

Since the beginning of the program, the goal of the NECAP has been to measure what students know and are able to do by using a variety of test item types; the program was structured to use both common and matrix-sampled items. (Common items are those taken by all students at a given grade level; matrix-sampled items make up a pool of items that is divided among the multiple forms of the test at each grade level.) This design provides reliable and valid results at the student level and breadth of coverage of a content area for school results while minimizing testing time.

Embedded Field Test

The NECAP includes an embedded field test in all content areas except writing. Because the field test is taken by all students, it provides the sample needed to produce reliable data with which to inform the process of selecting items for future tests.

Embedding the field test achieves two other objectives. First, it creates a pool of replacement items needed due to natural attrition caused by the release of common items each year in reading and mathematics. Second, embedding field-test items into the operational test ensures that students take the items under operational conditions.

TEST BOOKLET DESIGN

To accommodate the embedded field test in the 2005–06 NECAP, there were nine unique test forms at each grade. In all reading and mathematics test sessions, the field-test items were distributed among the common items in a way that was not evident to test takers. The writing design called for one common test form that was made up of a single writing prompt, four CR items, and ten MC items.

READING TEST DESIGN

Table 2-1 summarizes the numbers and types of items that were used in the NECAP reading assessment for 2005–06. Each MC item was worth one point, and each CR item was worth four points.

Table 2-1
NECAP Reading Numbers of Items and Types

Common – 2 long and 2 short passages plus 4 stand-alone MC		Matrix – Equating Forms 1,2,3 1 long and 1 short passages plus 2 stand-alone MC		Matrix – FT Forms 4-7 1 long and 1 short passages plus 2 stand-alone MC		Matrix – FT Forms 8–9 3 short passages plus 2 stand-alone MC		Total per student – 3 long and 3 short or 2 long and 5 short passages plus 6 stand-alone MC	
MC	CR	MC	CR	MC	CR	MC	CR	MC	CR
28	6	14	3	14	3	14	3	42	9

READING BLUEPRINT

As indicated earlier, the assessment framework for reading was based on the *NECAP Grade Level Expectations*, and all items on the NECAP test were designed to measure a specific GLE. The reading passages on the NECAP test are broken down into the following categories:

- **Literary passages** presenting a variety of forms: modern narratives; diary entries; drama; poetry; biographies; essays; excerpts from novels; short stories; and traditional narratives, such as fables, tall tales, myths, and folktales.
- **Informational passages** that are factual texts and often deal with the areas of science and social studies. These passages are taken from such sources as newspapers, magazines, and excerpts from books. Informational text also includes directions, manuals, or recipes.

  The passages are authentic texts—selected from grade-level-appropriate reading sources—that students would be likely to experience in both the classroom and independent reading. Passage is written specifically for the assessment; all are collected from published works.

The items on the NECAP test are categorized by both the type of passage associated with the item and whether the item measured lower or higher comprehension. The level of comprehension is designated as either "Initial Understanding" or "Analysis and Interpretation." Word identification and vocabulary skills are assessed, primarily through MC items, at each grade level. The distribution of emphasis for reading is shown in Table 2-2.

Table 2-2
Reading Distribution of Emphasis

	2 (3)	3 (4)	4 (5)	5 (6)	6 (7)	7 (8)
	Target	Target	Target	Target	Target	Target
Word Identification Sills and Strategies	20%	15%	0%	0%	0%	0%
Vocabulary Strategies/Breadth of Vocabulary	20%	20%	20%	20%	20%	20%
Initial Understanding of Literary Text	20%	20%	20%	20%	15%	15%
Initial Understanding of Informational Text	20%	20%	20%	20%	20%	20%
Analysis and Interpretation of Literary Text	10%	15%	20%	20%	25%	25%
Analysis and Interpretation of Informational Text	10%	10%	20%	20%	20%	20%
	100%	100%	100%	100%	100%	100%

The subcategory reporting structure for reading is shown in Table 2-3. Also, displayed are the maximum possible number of raw score points that students could earn.

Table 2-3
NECAP Reading Reporting Subcategories and Possible Raw Score Points

Reading		Grade 3 Possible Points	Grade 4 Possible Points	Grade 5 Possible Points	Grade 6 Possible Points	Grade 7 Possible Points	Grade 8 Possible Points
Word ID/Vocabulary		20	20	10	10	10	10
Type of Text	Literary	16	16	21	21	21	22
	Informational	16	16	21	21	21	20
Level of Comprehension	Initial Understanding	20	20	26	21	17	18
	Analysis and Interpretation	12	12	16	21	25	24

With the exception of Word ID/Vocabulary items, reading items were reported in two ways: type of text and level of comprehension.

Table 2-4 lists the percentage of total score points assigned to each level of Depth of Knowledge in Reading.

Table 2-4
Reading Depth of Knowledge Percentages

Reading	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Level 1	34%	27%	15%	17%	15%	17%
Level 2	58%	65%	70%	58%	44%	52%
Level 3	8%	8%	15%	25%	41%	31%

MATHEMATICS TEST DESIGN

Table 2-5 summarizes the numbers and types of items that were used in the NECAP mathematics assessment in grades 3 and 4 for 2005–06. Table 2-6 summarizes the numbers and types of items that were used in the NECAP mathematics assessment in grades 5 through 8 for 2005–06. Each MC item was worth one point, SA items were worth one and two points, and each CR item was worth four points. The score points at each grade level were evenly divided so that the MC items represented approximately fifty percent of the possible score points and the SA and CR items together represented approximately fifty percent of the score points.

Table 2-5
NECAP Mathematics Numbers of Items and Types for Grades 3 and 4

Common			Matrix – Equating			Matrix – FT			Total per Student		
MC	SA1	SA2	MC	SA1	SA2	MC	SA1	SA2	MC	SA1	SA2
35	10	10	6	2	2	3	1	1	44	13	13

Table 2-6
NECAP Mathematics Numbers of Items and Types for Grades 5 through 8

Common				Matrix – Equating				Matrix – FT				Total per Student			
MC	SA1	SA2	CR	MC	SA1	SA2	CR	MC	SA1	SA2	CR	MC	SA1	SA2	CR
32	6	6	4	6	2	2	1	3	1	1	1	41	9	9	6

THE USE OF CALCULATORS ON THE NECAP

The mathematics specialists from the NH, RI, and VT Departments of Education who designed the mathematics assessment acknowledge the importance of mastering arithmetic algorithms. At the same time, they understand that the use of calculators is a necessary and important skill. Calculators can save time and prevent error in the measurement of some higher-order thinking skills and allow students to work more sophisticated and intricate problems. For these reasons, it was decided that calculators should be permitted in Session 1 of the NECAP mathematics assessment and prohibited in Sessions 2 and 3.

MATHEMATICS BLUEPRINT

The assessment framework for mathematics was based on the *NECAP Grade Level Expectations*, and all items on the NECAP test were designed to measure a specific GLE. The mathematics items are organized into four content standards as shown on the following list:

- **Numbers and Operations:** Students understand and demonstrate a sense of what numbers mean and how they are used. Students understand and demonstrate computation skills.

- **Geometry and Measurement:** Students understand and apply concepts from geometry. Students understand and demonstrate measurement skills.
- **Functions and Algebra:** Students understand that mathematics is the science of patterns, relationships, and functions. Students understand and apply algebraic concepts.
- **Data, Statistics, and Probability:** Students understand and apply concepts of data analysis. Students understand and apply concepts of probability.

In addition, problem solving, reasoning, connections, and communication are embedded throughout the GLEs. The distribution of emphasis for Mathematics is shown in Table 2-7.

Table 2-7
Mathematics Distribution of Emphasis

	2 (3)	3 (4)	4 (5)	5 (6)	6 (7)	7 (8)
Numbers and Operations	55%	50%	45%	40%	30%	20%
Geometry and Measurement	15%	20%	20%	25%	25%	25%
Functions and Algebra	15%	15%	20%	20%	30%	40%
Data, Statistics, and Probability	15%	15%	15%	15%	15%	15%
	100%	100%	100%	100%	100%	100%

As shown in Table 2-8, the goal for distribution of score points, or balance of representation across the four content strands, varies from grade to grade.

Table 2-8
NECAP Mathematics Reporting Subcategories and Possible Raw Score Points

Mathematics	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	Possible Points	Possible Points	Possible Points	Possible Points	Possible Points	Possible Points
Numbers and Operations	35	32	30	26	20	13
Geometry and Measurement	10	13	13	17	16	16
Functions and Algebra	10	10	13	13	19	27
Data, Statistics, and Probability	10	10	10	10	11	10

Table 2-9 lists the percentage of total score points assigned to each level of Depth of Knowledge in mathematics.

Table 2-9
Mathematics Depth of Knowledge Percentages

Mathematics	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Level 1	29%	24%	20%	17%	24%	20%
Level 2	63%	62%	63%	70%	59%	62%
Level 3	8%	14%	17%	13%	17%	18%

WRITING TEST DESIGN

Table 2-10 summarizes the numbers and types of items that were used in the NECAP mathematics assessment in grades 5 and 8 for 2005–06. Each MC item was worth one point, each CR item was worth four points, each SA item was worth one point, and the writing prompt was worth 12 points.

Table 2-10
NECAP Writing Numbers of Items and Types for Grades 5 and 8

All Common – Total Per Student			
MC	CR	SA	WP
10	3	3	1

WRITING BLUEPRINT

The assessment framework for writing was based on the *NECAP Grade Level Expectations*, and all items on the NECAP test were designed to measure a specific GLE. The content standards in writing identify four major genres that are assessed in the writing portion of the NECAP test each year.

- **Writing in response to literary text**
- **Writing in response to informational text**
- **Narratives**
- **Informational writing (report/procedure for Grade 5 and persuasive at Grade 8)**

The writing prompt and the three CR items each address a different genre. In addition, structures and conventions of language are assessed through multiple-choice items and throughout the student's writing. The prompts and constructed-response items were developed with the following criteria as guidelines:

- The prompts must be interesting to students
- the prompts must be accessible to all students (i.e., all students would have something to say about the topics) and
- the prompts must generate sufficient text to be effectively scored

The subcategory reporting structure for writing is shown in Table 2-11. Also displayed are the maximum possible number of raw score points that students could earn.

Table 2-11
NECAP Writing Reporting Subcategories and Possible Raw Score Points

Writing	Grade 5	Grade 8
	Possible Points	Possible Points
Structures of Language and Writing Conventions	10	10
Short Responses	12	12
Extended Response	15	15

The subcategory "Short Responses" lists the total raw score points from the three CR items, where as the subcategory "Extended Response" lists the total raw score points from the three short-answer items and the writing prompt.

Table 2-12 lists the percentage of total score points assigned to each level of Depth of Knowledge in writing.

Table 2-12
Writing Depth of Knowledge Percentages

Writing	Grade 5	Grade 8
Level 1	19%	22%
Level 2	41%	38%
Level 3	40%	40%

TEST SESSIONS

The NECAP tests were administered to grades 3 through 8 during October 3–26 in October 2005. Schools were able to schedule testing sessions at any time during the two weeks of this period, provided they followed the sequence in the scheduling guidelines detailed in test administration manuals and that all testing classes within a school were on the same schedule. The third week was reserved for make-up testing of students who were absent from initial test sessions.

The timing and scheduling guidelines for the NECAP tests were based on estimates of the time it would take an average student to respond to each type of item that makes up the test:

- multiple-choice – 1 minute
- short-answer (1 point) – 1 minute
- short-answer (2 point) – 2 minutes
- constructed-response – 10 minutes
- long writing prompt – 45 minutes

For the reading tests, the scheduling guidelines included an estimate of 10 minutes to read the stimulus material used in the assessment. Tables 2-13 through 2-16 show the distribution of items across the test sessions for each content area for each grade level.

Table 2-13
NECAP Reading Test Sessions for Grades 3 through 8

	Session 1	Session 2	Session 3
	1 long and 1 short passage plus 2 stand-alone MC	1 long and 1 short passage plus 2 stand-alone MC	1 long and 1 short passage plus 2 stand-alone MC
MC	14	14	14
CR	3	3	3

Table 2-14
NECAP Mathematics Test Sessions for Grades 3 and 4

	Session 1	Session 2	Session 3
MC	15	15	14
SA1	4	3	6
SA2	4	5	4

Table 2-15
NECAP Mathematics Test Sessions for Grades 5 through 8

	Session 1	Session 2	Session 3
MC	14	14	13
SA1	3	3	3
SA2	3	3	3
CR	2	2	2

Table 2-16
NECAP Writing Test Sessions for Grades 5 and 8

	Session 1	Session 2
MC	10	0
CR	3	0
SA	0	3
WP	0	1

Though the guidelines for scheduling are based on the assumption that most students will complete the test within the time estimated, each test session was scheduled so that additional time was provided for students who needed it. Up to one hundred percent additional time was allocated for each session (e.g., a 50-minute session could have up to an additional 50 minutes).

If classroom space was not available for students who required additional time to complete the tests, schools were allowed to consider using another space, such as the guidance office, for this purpose. If additional areas were not available, it was recommended that each classroom being used for

test administration be scheduled for the maximum amount of time. Detailed instructions on test administration and scheduling were provided in the test coordinators' and administrators' manuals.

2.5 ACCESSIBILITY

A major area of emphasis in both the NECAP *Request for Pro*posals and on most subsequent planning and management meeting agendas was the importance of making the NECAP program as accessible as possible to as many students as possible. Activities to address this issue were focused around five areas: training, publication specifications, focused reviews, accommodations, and analysis of test data. Details of activities on each of these areas are listed below and on the following pages:

Training

- Before beginning work, Measured Progress provided two different training sessions on the principles of universal design for test development, publications, and program management staff. These sessions were conducted by the Center for Applied Special Technology (CAST) and the special-education division of Measured Progress.
- The special-education division of Measured Progress also provided training on the principles of universal design to all IRC members at the beginning of their work.

Publication Specifications

- A style guide for NECAP publications was developed consistent with guidelines proffered by the National Center on Educational Outcomes (NCEO). In a few cases, exceptions were made. For example, the NCEO recommended a specific number of words in a line of text within a reading passage. Though this worked well with certain font sizes, in some cases the specification was changed for a particular grade/developmental level.
- "Think Bubbles" that were added to some items contained information a student could use to answer a question but was extraneous to the actual construct being measured by the item.

- Schools who requested a Braille form of the test were contacted so that each student was provided with either a contracted or uncontracted version, depending on what version was being used in a student's program of studies.
- In keeping with a change in the recommended font size for large-print test booklets by the American Printing House for the Blind, the size of the font was changed from 16 point to 20 point.
- Large numbers of graphics were included in the test so as to maximize students' ability to interact with items that required spatial visualization. Manipulatives were also provided for some items for this same purpose.

Focused Reviews

- As stated, IRCs received some training on universal design before beginning their review. In addition, committee members responded to a series of questions for each item, two of which were "Is the item developmentally appropriate?" and "Is the item as accessible as possible?" Committee members' responses to these questions were used to revise items when necessary. The composition of IRC membership included special-education and English language learner (ELL) teachers on each committee whenever possible.
- A separate bias/sensitivity review committee was formed to examine any issues in this area that might prohibit students from maximizing the quality of their response to an item. Once again, committees were trained in the principles of universal design, and their feedback was included in the revision process.
- The NCEO conducted an independent review of the items, and their findings were provided to the states and Measured Progress test developers for use in revision and future item development.

- The NECAP Technical Advisory Committee includes a representative from the NCEO to ensure that discussions of technical issues included the perspective of providing maximum accessibility.

DIF Analyses

- In addition to more traditional Differential Item Functioning (DIF) analyses (e.g., male-female and black-white), a series of DIF analyses were conducted to examine relationships among a variety of subgroups. For example, analyses were conducted on special education against non–special education, learning disabled against non–learning disabled, and ELL against non-ELL, and on each state's population against each other state. Again, results of these analyses were used in the item revision/selection process.

Accommodations

- There was a recognition that, though every effort possible was being made to provide a test that would be as accessible as possible, a need still remained to allow some students to take the test with accommodations. An operating principle employed during the development of the accommodations protocols and policy development was to allow only accommodations that would not change the construct of what was being measured by the item. A complete description of the accommodations allowed and the process for schools to use when employing the can be found in *NECAP Accommodations, Guidelines, and Procedures: Administrator Training Guide.*

More specific details about these activities can be found in the pertinent sections of this technical report.

CHAPTER 3—TEST ADMINISTRATION

3.1 RESPONSIBILITY FOR ADMINISTRATION

As indicated in the *Principal/Test Coordinator Manual*, principals and/or their designated NECAP test coordinator were responsible for the proper administration of the NECAP. Manuals containing explicit directions and scripts for test administrators to read aloud to students were used to ensure the uniformity of administration procedures from school to school.

3.2 ADMINISTRATION PROCEDURES

Principals and/or their school's designated NECAP coordinator were instructed to read the *Principal/Test Coordinator Manual* before testing and to be familiar with the instructions provided in the *Test Administrator Manual*. The *Principal/Test Coordinator Manual* provided each school with checklists to help them to prepare for testing. The checklists outlined tasks for school staff to perform before, during, and after test administration. Along with these checklists, the *Principal/Test Coordinator Manual* outlined the nature of the testing material being sent to each school, how to inventory the material, how to track it during administration, and how to return the material after testing was complete. The *Test Administrator Manual* also included checklists for the administrators to prepare themselves, their classrooms, and the students for the administration of the test. The *Test Administrator Manual* contained sections that detailed the procedures to be followed for each test session, and it contained instructions on preparing the material before giving it to the principal/test coordinator for its return to Measured Progress.

3.3 PARTICIPATION REQUIREMENTS AND DOCUMENTATION

The legislation's intent is for **all** students in grades 3 though 8 to participate in the NECAP through standard administration, administration with accommodations, or alternate assessment. Furthermore, any student who is absent during any session of the NECAP is expected to take a makeup test within the three-week testing window.

Schools were required to return a student answer booklet for every enrolled student in the grade level. On those occasions when it was deemed impossible to test a particular student, school personnel were required to inform their Department of Education. The states included a grid on the student answer booklets that listed the approved reasons why a student answer booklet could be returned blank for one or more sessions of the test.

- **Student completed the Alternate Assessment for the 2004–2005 school year**
  If a student completed the alternate assessment in the previous school year, the student was not required to participate in the NECAP in 2005-06.

- **Student is new to the United States after October 1, 2004 and is LEP (reading and writing only)**
  First-year LEP students that took the ACCESS test of English language proficiency as scheduled in their states were not required to take the reading and writing tests in 2005–06. However, these students were required to take the mathematics test in 2005–06.

- **Student withdrew from school after October 1, 2005**
  If a student withdrew after October 1, 2005 but before to completing all of the test sessions, school personnel were instructed to code this reason on the student's answer booklet.

- **Student enrolled in school after October 1, 2005**

  If a student enrolled after October 1, 2005 and was unable to complete all of the test sessions before to the end of the testing administration window, school personnel were instructed to code this reason on the student's answer booklet.

- **State-approved special consideration**

  Each state department of education had a process for documenting and approving circumstances that made it impossible or not advisable for a student to participate in testing. Schools were required to obtain state approval before beginning testing.

- **Student was enrolled in school on October 1, 2005 and did not complete test for reasons other than those listed above**

  If a student was not tested for a reason not stated above, school personnel were instructed to code this reason on the student's answer booklet. These "Other" categories were considered "not state-approved."

Tables 3-1, 3-2, and 3-3 list the participation rates of the three states combined in reading, writing, and mathematics.

Table 3-1
Participation Rates for NECAP Reading

Category	Description	Enrollment	Not Tested State-Approved	Not Tested Other	Tested	Percent
All	All Students	210075	4401	1008	204666	97
gender	Male	107972	2565	599	104808	97
gender	Female	101343	1658	384	99301	98
gender	Not Reported	760	178	25	557	73
ethnic	AIAN	822	21	5	796	97
ethnic	Asian	4683	151	19	4513	96
ethnic	Black	8399	244	50	8105	96
ethnic	Hispanic	15182	567	62	14553	96
ethnic	NHPI	461	26	11	424	92
ethnic	White	179458	3292	834	175332	98
ethnic	Not Reported	1070	100	27	943	88
lep	Current	6279	735	38	5506	88
lep	Monitoring Year 1	1169	16	3	1150	98
lep	Monitoring Year 2	841	4	5	832	99
lep	Other	201786	3646	962	197178	98
iep	iep	33275	2946	302	30027	90
iep	Other	176800	1455	706	174639	99
ses	ses	59350	2262	325	56763	96
ses	Other	150725	2139	683	147903	98
migrant	migrant	248	54	1	193	78
migrant	Other	209827	4347	1007	204473	97
Title 1	Title 1	26084	545	73	25466	98
Title 1	Other	183991	3856	935	179200	97
Plan 504	Plan 504	62	3	0	59	95
Plan 504	Other	210013	4398	1008	204607	97

Table 3-2
Participation Rates for NECAP Mathematics

Category	Description	Enrollment	Not Tested State Approved	Not Tested Other	Tested	Percent
All	All Students	210075	3572	1137	205366	98
gender	Male	107972	2054	678	105240	97
gender	Female	101343	1384	426	99533	98
gender	Not Reported	760	134	33	593	78
ethnic	AIAN	822	17	6	799	97
ethnic	Asian	4683	46	22	4615	99
ethnic	Black	8399	148	50	8201	98
ethnic	Hispanic	15182	205	85	14892	98
ethnic	NHPI	461	24	13	424	92
ethnic	White	179458	3038	933	175487	98
ethnic	Not Reported	1070	94	28	948	89
lep	Current	6279	76	33	6170	98
lep	Monitoring Year 1	1169	2	4	1163	99
lep	Monitoring Year 2	841	1	6	834	99
lep	Other	201786	3493	1094	197199	98
iep	iep	33275	2752	348	30175	91
iep	Other	176800	820	789	175191	99
ses	ses	59350	1699	388	57263	96
ses	Other	150725	1873	749	148103	98
migrant	migrant	248	30	1	217	88
migrant	Other	209827	3542	1136	205149	98
Title 1	Title 1	23221	274	72	22875	99
Title 1	Other	186854	3298	1065	182491	98
Plan 504	Plan 504	62	1	0	61	98
Plan 504	Other	210013	3571	1137	205305	98

Table 3-3
Participation Rates for NECAP Writing

Category	Description	Enrollment	Not Tested State Approved	Not Tested Other	Tested	Percent
All	All Students	71817	1373	586	69858	97
gender	Male	36707	786	379	35542	97
gender	Female	34848	534	197	34117	98
gender	Not Reported	262	53	10	199	76
ethnic	AIAN	290	6	5	279	96
ethnic	Asian	1516	46	11	1459	96
ethnic	Black	2729	86	36	2607	96
ethnic	Hispanic	4908	182	53	4673	95
ethnic	NHPI	181	10	12	159	88
ethnic	White	61888	1022	457	60409	98
ethnic	Not Reported	305	21	12	272	89
lep	Current	1890	235	24	1631	86
lep	Monitoring Year 1	369	2	2	365	99
lep	Monitoring Year 2	217	0	2	215	99
lep	Other	69341	1136	558	67647	98
iep	iep	11782	914	226	10642	90
iep	Other	60035	459	360	59216	99
ses	ses	19498	702	241	18555	95
ses	Other	52319	671	345	51303	98
migrant	migrant	74	10	1	63	85
migrant	Other	71743	1363	585	69795	97
Title 1	Title 1	7018	176	24	6818	97
Title 1	Other	64799	1197	562	63040	97
Plan 504	Plan 504	26	0	0	26	100
Plan 504	Other	71791	1373	586	69832	97

3.4 ADMINISTRATOR TRAINING

In addition to distributing the *Principal/Test Coordinator* and *Test Administrator Manual*s, the NH, RI, and VT Departments of Education, along with Measured Progress, conducted test administration workshops in five separate regional locations in each state to inform school personnel about the NECAP and to provide training on the policies and procedures regarding administration of the NECAP tests.

3.5 DOCUMENTATION OF ACCOMMODATIONS

The *Principal/Test Coordinator* and *Test Administrator Manual* provided directions for coding the information related to accommodations and modifications on page 2 of the student answer booklet. All accommodations used during any test session were required to be coded in by authorized school personnel, not students, after testing was completed.

An *Accommodations, Guidelines, and Procedures: Administrator Training Guide* was also produced to provide detailed information on planning and implementing accommodations. This guide can be located on each state's Department of Education Web site. The states collectively made the decision that accommodations were available to all students on the basis of individual need regardless of disability status. Decisions regarding accommodations were to be made by the students' educational team on an individual basis and were to be consistent with those used during the students' regular classroom instruction. Making accommodations decisions based on an entire-group basis rather than on an individual basis was not permitted. If the decision made by a student's educational team required an accommodation not listed in the state-approved Table of Standard Test Accommodations, schools were instructed to contact the Department of Education in advance of testing for specific instructions for coding the "Other Accommodations (E)" and/or "Modifications (F)" section.

Tables 3-4 and 3-5 show the accommodations observed for the October 2005 NECAP administration. The Table of Standard Test Accommodations can be found in Appendix B.

Table 3-4
NECAP Accommodation Frequencies for NECAP 2005: Grades 3 through 5

	Grade 3		Grade 4		Grade 5		
Accom.	Reading	Math	Reading	Math	Reading	Math	Writing
A01	721	710	616	607	574	582	539
A02	3714	3759	3783	3916	3858	3971	3735
A03	1357	1318	1147	1162	1250	1243	1190
A04	288	299	252	252	246	245	226
A05	23	20	8	7	7	8	9
A06	22	16	57	53	16	18	17
A07	1688	1684	1755	1760	1826	1878	1773
A08	1138	1179	1161	1192	1064	1082	972
A09	4	6	1	2	5	6	6
B01	219	211	244	243	254	212	183
B02	1760	1692	1746	1695	1833	1829	1706
B03	2240	2144	2286	2368	2434	2734	2242
C01	4	4	4	4	0	0	0
C02	55	52	38	39	30	27	25
C03	26	29	18	14	28	29	27
C04	0	3325	0	2985	0	2861	2539
C05	480	385	346	267	333	286	258
C06	86	102	59	74	29	84	30
C07	614	570	617	510	588	544	484
C08	14	6	8	8	2	5	3
C09	242	192	201	145	147	112	105
C10	17	7	33	13	14	13	13
C11	54	50	58	48	34	31	27
C12	0	23	0	31	0	26	13
C13	0	3	0	3	0	0	0
D01	25	13	29	17	75	47	137
D02	103	88	91	80	100	72	100
D03	17	11	6	7	17	17	15
D04	136	128	116	115	161	157	123
D05	844	747	731	628	600	486	0
D06	18	10	9	10	14	12	0
E01	14	8	10	10	4	34	5
E02	0	0	0	0	0	0	53
F01	0	50	0	30	0	39	0
F02	42	0	19	0	24	0	0
F03	8	3	4	2	2	2	4

Table 3-5
Accommodation Frequencies for NECAP 2005: Grades 6 through 8

	Grade 6		Grade 7		Grade 8		
Accom.	Reading	Math	Reading	Math	Reading	Math	Writing
A01	453	459	348	332	294	280	257
A02	3603	3553	3393	3352	3273	3278	3171
A03	826	804	529	493	631	613	556
A04	195	187	239	220	232	231	223
A05	10	9	9	9	25	16	13
A06	14	12	9	5	19	16	14
A07	1858	1747	1626	1573	1522	1498	1470
A08	653	644	415	417	384	365	339
A09	16	13	4	4	18	15	13
B01	198	196	140	135	150	143	121
B02	1433	1412	1060	1044	1016	986	922
B03	2232	2317	2019	2066	1732	1729	1589
C01	0	0	1	1	1	1	1
C02	35	31	25	24	23	21	20
C03	13	13	8	5	14	15	13
C04	0	1886	0	1524	0	1211	1187
C05	142	130	83	61	92	72	70
C06	21	36	40	47	57	56	38
C07	302	305	241	228	214	230	224
C08	8	12	14	13	3	4	3
C09	52	37	10	6	13	12	8
C10	9	2	1	2	7	5	5
C11	19	20	12	10	17	14	14
C12	0	38	0	103	0	82	71
C13	0	0	0	1	0	0	0
D01	126	74	154	74	160	87	215
D02	60	50	42	32	36	25	33
D03	7	11	12	10	25	4	6
D04	98	99	81	78	64	37	30
D05	367	294	238	191	159	163	0
D06	10	7	7	5	6	6	0
E01	2	10	12	11	1	2	7
E02	0	0	0	0	0	0	30
F01	0	41	0	60	0	61	0
F02	18	0	32	0	11	0	0
F03	1	1	1	0	0	0	1

3.6 TEST SECURITY

Maintaining test security is critical to the success of the New England Common Assessment program and the continued partnership among the three states. The *Principal/Test Coordinator Manual* and the *Test Administrator Manual*s explain in detail all of the test security measures and test administration procedures that are to be adhered to by the schools with respect to the handling of the secure test materials and the administering of the tests. School personnel were informed that any concerns about breaches in test security were to be reported to the schools' test coordinator and principal immediately. The test coordinator and/or principal was responsible for immediately reporting the concern to the district superintendent and the state director of assessment at the department of education. Test Security was also strongly emphasized at test administration workshops that were conducted in all three states.

The three states also required the principal of each school that participated in testing to log on to a secure website to complete the *Principal's Certification of Proper Test Administration* form for each grade level tested. Principal's were requested to provide the number of secure tests received from Measured Progress, the number of tests administered to students, and the number of secure test materials that they were returning to Measured Progress. Principal's were then instructed to print off a hard copy of the form, sign it, and return it with their test materials shipment. By signing the form, the principal was certifying that the tests were administered according to the test administration procedures outlined in the *Principal/Test Coordinator* and *Test Administrator* Manuals, that they maintained the security of the tests, that no secure material was duplicated or in anyway retained in the school, and that all test materials have been accounted for and returned to Measured Progress.

3.7 TEST AND ADMINISTRATION IRREGULARITIES

During the first few days of test administration, a printing error was discovered in approximately twenty grade 6, form 6 NECAP test booklets. Four schools called the NECAP Service Center or their state Department of Education and reported that pages six through fifteen were missing from one or more of their grade 6 test booklets. The print vendor determined that the error occurred during the setup for the form 6 booklets and estimated that only approximately twenty test booklets should be defective. The vendor explained that calipers are used to measure the thickness of each test booklet. The calipers are placed after it is determined that all pages in a booklet are in order and accounted for. The binding machine will shut down if pages are missing; therefore, the defective booklets were used for setup before the calipers were in place. Not all the booklets used to set up the machine are complete booklets and should be thrown away after setup. The vendor believed that the grade 6, form 6 test booklets were not thrown away and were mistakenly sent to Measured Progress. In total, schools reported twenty defective booklets, confirming the vendor's belief that they were booklets used for setup. All affected schools either replaced the defective booklets with extra grade 6 test booklets they already had in the school or Measured Progress immediately sent new booklets to the school. No NECAP report was affected by these irregularities.

3.8 TEST ADMINISTRATION WINDOW

The test administration window was October 3–25, 2005.

3.9 NECAP SERVICE CENTER

To provide additional support to schools before, during, and after testing, Measured Progress established the NECAP Service Center. The additional support that the Service Center provides is an element essential to the successful administration of any statewide assessment program. It provides a centralized location to which individuals in the field can call using a toll-free number to ask specific questions or to report any problems they may be experiencing.

The Service Center was staffed by representatives at varying levels based on need volume and was available from 8:00 AM to 4:00 PM beginning two weeks before to the start of testing and ending two weeks after testing. The representatives were responsible for receiving, responding to, and tracking calls and routing issues to the appropriate person(s) for resolution. All calls were logged into a database that was provided to each state after testing was completed.

CHAPTER 4—SCORING

4.1 IMAGING PROCESS

After the 2005–06 NECAP student answer booklets had been logged in, identified with appropriate scannable pre printed school information header sheets, examined for extraneous materials, and batched, they were moved into the scanning area for imaging. This area is the last stop in the processing loop in which the student answer booklets themselves are handled.

At this point, the student answer booklets were scanned, and the necessary information to produce the required reports was captured and converted into an electronic format, including all student identification and demographics, CR answers, and digital image clips of hand-written writing-prompt responses. The digital image clip information allowed Measured Progress to replicate student responses just as they appeared on the originals, so they can be transferred onto the readers' monitors for scoring. From this point on, the entire process—data processing, benchmarking, scoring, data analysis, and reporting—was accomplished without further reference to the originals.

The first step in this conversion was the removal of the booklet bindings so that the individual pages could pass through the scanners, one at a time. Once cut, the sheets were put back into their proper boxes and placed in storage until needed for the scanning/imaging process.

Customized scanning programs for all scannable documents were prepared to selectively read the student answer booklets and to format the scanned information electronically according to pre determined requirements. Any information (including MC response data) that had been designated time-critical or process-critical was handled first.

4.2 QUALITY CONTROL

The ScanQuest system is equipped with many built-in safeguards that prevent data errors. In addition to numerous real-time quality control checks, duplex read, and on-line editing capability, the I840 scanners offer features that make them compatible with Internet technology.

The scanning hardware is continually monitored for conditions that will cause the machine to shut down if standards are not met. It will display an error message and prevent further scanning until the condition is corrected. The areas monitored include document page and integrity checks and many internal checks of electronic functions.

Before every scanning shift begins, Measured Progress's operators perform a daily diagnostic routine. This is yet another step to protect data integrity and one that has been performed faithfully for the many years that we have been involved in production scanning. In the event that any inconsistencies are identified, the operator will calibrate that machine and perform the test again. If the machine is still not up to standard, we call for assistance from our field service engineer.

As a final safeguard, spot checks of scanned files, bubble by bubble and image by image, were routinely made throughout scanning runs to ensure the integrity of the data. After the data had been entered and the scanning logs and other paperwork completed, the booklets themselves were put into storage (where they stay for at least 180 days beyond the close of the fiscal year). When it had been determined that the databases were complete and accurate, they were exported for analysis and made available for many other processing options. Completed batches were loaded onto our local area network (LAN) for transfer to Measured Progress's proprietary iScore system for scoring. Those files were then used to identify (and print) papers to be used in the range-finding and standard-setting processes, and the data were made transferable via the Internet, CD-ROM, or optical disk.

4.3 HAND-SCORING

iSCORE

After the 2005–06 test material had been loaded into the LAN, qualified readers accessed electronically scanned images of student responses by sitting at computer terminals. The readers evaluated each response and recorded each student's score via keypad or mouse entry through the *iScore* system. The *iScore* system itself ensures the security of student responses and test items: All scoring is "blind" (i.e., no student names or raw scores are associated with viewed responses, and all scoring personnel are subject to the same nondisclosure requirements and supervision as are regular Measured Progress staff).

Readers had authorization to access only those student responses to only those items that they were qualified to score. When a reader finished one response, the next response appeared immediately on the computer screen. In that way, the system guaranteed the complete anonymity of individual students and ensured the randomization of responses during scoring.

Although *iScore* utilizes conventional scoring techniques, it offers numerous benefits, not the least of which is raising the bar on scoring process capability. Some of the benefits include

- real-time information on scorer reliability, read-behinds, and overall process monitoring;
- early access to subsets of data for such tasks as standard setting;
- reduced material handling, which not only saves time and labor but enhances the security of materials; and
- immediate access to samples of student responses and scores for reporting and analysis through electronic media.

SCORER QUALIFICATIONS

Under the Director of Scoring Services, scoring staff carried out the various scoring operations. The staff included

- chief readers (CRs), who oversaw all training and scoring within particular content areas;
- quality assurance coordinators (QACs), who led range finding and training activities and monitored scoring consistency and rates;
- senior readers (SRs), who performed read-behinds of readers and assisted at scoring tables as necessary; and
- readers, who performed the bulk of the scoring.

Table 4-1 summarizes the qualifications of the 2005-06 NECAP quality assurance coordinators and readers.

Table 4-1
NECAP Mathematics Test Sessions for Grades 5 through 8

October 2005 Administration					
Scoring Responsibility	Educational Credentials				Total
	Doctorate	Masters	Bachelors	Other	
QACs	4.9	30.4	56.9	7.8	100%
Readers	3.8	30.2	58.8	7.2	100%

BENCHMARKING

Before the scheduled start of scoring activities, scoring center staff reviewed test items and scoring guides for benchmarking. At that point, CRs and selected QACs prepared scorer training materials. Measured Progress's scoring staff (including test developers) selected one or two anchor examples for each item score point. An additional six to ten responses per item were chosen as part of the training pack. The anchor pack consisted of mid-range exemplars, where as the training pack exemplars illustrated the range within each score point. The CRs, who worked closely with QACs for each content area, facilitated the selection of response exemplars. One of the greatest difficulties in the

selection of anchor and training exemplars was finding a sufficient number of papers representing the highest scores, as such scores are fairly rare. All of the benchmarking materials initially selected by Measured Progress were reviewed by the content representatives from each state. Based on their recommendations, the anchor exemplars and training packs were modified, finalized, and approved for scorer training.

Selecting and Training Quality Assurance Coordinators and Senior Readers

Because the read-behinds performed by the QACs and SRs moderated the scoring process and thus maintained the integrity of the scores, individuals to fill those positions were selected for their accuracy. In addition, QACs, who train readers to score each item in their content areas, were selected for their ability to instruct and for their level of expertise in their content areas. For this reason, QACs typically are retired teachers who have demonstrated a high level of expertise in their respective disciplines. The ratio of QACs and SRs to readers was approximately 1:11.

Selecting and Training Readers

Applicants were required to demonstrate their ability by participating in a preliminary scoring evaluation. The *iScore* system enables Measured Progress to efficiently measure a prospective reader's ability to score student responses accurately. After having participated in a training session, applicants were required to achieve at least eighty percent exact scoring agreement for a qualifying pack consisting of ten responses to a predetermined item in their content area (twenty responses for equating items). Those ten responses were randomly selected from a bank of approximately 150, all of which had been selected by QACs and approved by the CRs, developers, and content representatives from each state.

The QACs first applied the language of the scoring guide for an item to its anchor pack exemplars. Once discussion of the anchor pack had concluded, readers scored the training pack exemplars. The QACs then reviewed the training pack scoring by the readers and answered any

questions that readers had before actual scoring began. With this system, two aspects of scoring efficiency are in conflict. First, in order to minimize training expense, it is desirable to train each reader on as few items as possible. Second, to prevent reader drift and to minimize retraining requirements, it is desirable to score a given item in a brief period of time. However, the lower the number of unique items each reader scores, the greater the number of readers required to score that item quickly. To minimize that conflict, the readers for each content area were divided into two or more groups. On the first day of scoring, each group was trained to score a different item. When a group had completed all of an item's responses, those readers were trained on another item (or set).

MONITORING READERS

After a reader scored a student response, *iScore* determined whether that response should also be scored by another reader, scored by a QAC or SR, or routed for special attention. QACs and SRs used *iScore* to produce daily reader accuracy and speed reports. QACs and SRs were able to obtain current reader accuracy reports and speed reports on-line at any time.

The weighted averages of exact (both readers assigned the paper the same score), adjacent (the two readers scores differed by one point), and total (exact or adjacent) percent agreement are reported in Table 4-2. The weighting was based on the number of responses that were rescored for each question. (Note; These data underestimate scorer accuracy.) Blanks were included in both read-behind and double-blind scoring. Readers were instructed to score any questions for which the student had made a mark of any kind as a zero. However, in many instances it was impossible for the reader to tell whether a mark on the page was written by the student or whether there was a crease in the paper, bleed-through from the other side of the page, or dust on the image screen. In such instances, these responses were counted as neither exact nor adjacent agreement, though the effect of blanks and zeroes on student scores was identical.

Table 4-2
Scoring Consistency and Reliability
Double-Blind

Grade	Math			Reading			Writing		
	Exact	Adjacent	Total	Exact	Adjacent	Total	Exact	Adjacent	Total
3	95.3	4.50	99.8	70.3	26.3	96.6			
4	93.9	5.90	99.8	72.7	24.3	97.0			
5	89.0	10.20	99.2	65.8	32.4	98.2	68.5	29.8	98.3
6	89.8	9.40	99.2	62.8	35.1	97.9			
7	87.7	11.1	98.8	59.1	37.4	96.5			
8	91.1	8.30	99.4	63.3	34.3	97.6	60.1	36.4	96.5

Read-Behind

Grade	Math			Reading			Writing		
	Exact	Adjacent	Total	Exact	Adjacent	Total	Exact	Adjacent	Total
3	93.5	6.30	99.8	70.4	26.5	96.9			
4	90.0	9.80	99.8	72.2	25.2	97.4			
5	87.7	11.4	99.1	65.3	33.2	98.5	69.6	28.8	98.4
6	85.4	13.6	99.0	63.2	34.7	97.9			
7	83.6	14.4	98.0	61.8	35.9	97.7			
8	84.3	14.0	98.3	64.4	33.8	98.2	65.4	33.2	98.6

SCORING ACTIVITIES

Student response booklets were digitally scanned and scored on a file server for a dedicated, secure LAN. *iScore* then distributed digital images of student responses to readers. Training and scoring took place over a period of approximately three weeks. Items were randomly assigned to readers; thus, each item in a student's response booklet was more than likely scored by a different reader. By using the maximum possible number of readers for each student, the procedure effectively minimized error variance due to reader sampling. All common and matrix CR items in reading and mathematics were scored once with a two-percent read-behind to ensure consistency among readers and accuracy of individual readers. At grades 5 and 8, the common writing prompt was scored independently by two readers with the requirement that the two scores for each writing component had to be adjacent. Non adjacent scores were arbitrated. The combined scores given by the two readers resulted in the student's raw score on the writing prompt. Each of the three writing CR items was scored once with a two-percent read-behind, and these points were added to the points earned on the writing prompt and the points

earned on the ten MC items covering the structures of language and conventions, resulting in the total raw score for writing.

SCORING LOCATIONS

All of the oversight and administrative controls applied to the *iScore* database were managed for scoring at Measured Progress headquarters in Dover, NH. However, student responses were scored in three locations: Dover, NH; Troy, NY; and Aurora, CO. Table 4-3 shows the locations where each content area/grade level combinations were scored. It is important to note that no single item was scored in more than one location.

The *iScore* system monitored accuracy, reliability, and consistency across all scoring locations. Constant communication and coordination were accomplished through e-mail, telephone, faxes, and secure Web sites, to ensure critical information and scoring modifications were shared/implemented across all scoring locations.

Table 4-3
Content Area/Grade Level Scoring Locations

Content Area/Grade Level	Dover, NH	Troy, NY	Aurora, CO
Reading Grade 3	X		
Reading Grade 4	X		X
Reading Grade 5	X	X	X
Reading Grade 6	X	X	X
Reading Grade 7	X	X	X
Reading Grade 8	X	X	X
Mathematics Grade 3			X
Mathematics Grade 4		X	
Mathematics Grade 5			X
Mathematics Grade 6		X	
Mathematics Grade 7			X
Mathematics Grade 8	X	X	
Writing Grade 5	X		X

EXTERNAL OBSERVATIONS

All scoring locations were visited by at least one representative from each of the three Departments of Education during scoring. State assessment directors and content specialists from the three states were present at some point in each location during benchmarking, training, and live scoring throughout the scoring window. The state assessment directors and content specialists from the three states met with program management and scoring management staff from Measured Progress to share their observations and provide feedback. Recommendations that were a result of that meeting will be applied to the next round of scoring in 2006–07.

CHAPTER 5—SCALING AND EQUATING

5.1 ITEM RESPONSE THEORY SCALING

All NECAP items were calibrated using Item Response Theory (IRT). IRT uses mathematical models to define a relationship between an unobserved measure of student performance, usually referred to as *theta* (θ), and the probability (*p*) of getting a dichotomous item correct or of getting a particular score on a polytomous item. In IRT, it is assumed that all items are independent measures of the same construct (i.e., the same θ). Another way to think of θ is to consider it as a mathematical representation of the latent trait or construct of interest. Several common IRT models are used to specify the relationship between θ and *p* (Hambleton and van der Linden, 1997; Hambleton and Swaminathan, 1985). The process of determining the specific mathematical relationship between θ and *p* is called *item calibration*. After items are calibrated, they are defined by a set of parameters that specify a nonlinear, monotonically increasing relationship between θ and *p*. Once the item parameters are known, the $\hat{\theta}$ for each student can be calculated. In IRT, $\hat{\theta}$ is considered to be an estimate of the student's true score or a general representation of student performance and has some characteristics that may make its use preferable to the use of raw scores in equating.

For NECAP 2005, the three-parameter logistic (3PL) model and the graded-response model (GRM) were used for dichotomous and polytomous items, respectively. The 3PL model for dichotomous items can be defined as:

$$P_i\left(1\middle|\theta_j\right) = c_i + \left(1-c_i\right)\frac{\exp Da_i\left(\theta_j - b_i\right)}{1+\exp Da_i\left(\theta_j - b_i\right)}$$

where *i* indexes the items,

j indexes students,

a represents item discrimination,

b represents item difficulty,

c is the pseudo-guessing parameter, and

D is a normalizing constant equal to approximately 1.701.

In the GRM for polytomous items, an item is scored in k+1 graded categories that can be viewed as a set of k dichotomies. At each point of dichotomization (i.e., at each threshold), a two-parameter model can be used. This implies that a polytomous item with k+1 categories can be characterized by k item category threshold curves (ICTC) of the two-parameter logistic form:

$$P_{ik}^{*}\left(1\middle|\theta_j\right)=\frac{\exp Da_i\left(\theta_j-b_i+d_{ik}\right)}{1+\exp Da_i\left(\theta_j-b_i+d_{ik}\right)}$$

where *i* indexes the items,

j indexes students,

k indexes thresholds,

a represents item discrimination,

b represents item difficulty,

d represents a category step parameter, and

D is a normalizing constant equal to approximately 1.701.

After computing k item category threshold curves in the GRM, k+1 item category characteristic curves (ICCC) are derived by subtracting adjacent ICTC curves:

$$P_{ik}(1\mid\theta_j)=P_{i(k-1)}^{*}(1\mid\theta_j)-P_{ik}^{*}(1\mid\theta_j)$$

where P_{ik} represents the probability that the score on item *i* falls in category *k*, and

P_{ik}^{*} represents the probability that the score on item *i* falls above the threshold *k*

($P^*_{i0} = 1$ and $P^*_{i(k+1)} = 0$).

The GRM is also commonly expressed as:

$$P_{ik}\left(k\middle|\theta_j,\xi_i\right) = \frac{\exp\left[Da_i\left(\theta_j - b_i + d_k\right)\right]}{1+\exp\left[Da_i\left(\theta_j - b_i + d_k\right)\right]} - \frac{\exp\left[Da_i\left(\theta_j - b_i + d_{k+1}\right)\right]}{1+\exp\left[Da_i\left(\theta_j - b_i + d_{k+1}\right)\right]}$$

where ξ_i represents the set of item parameters for item *i*.

Finally, the ICC for polytomous items is computed as a weighted sum of ICCCs, where each ICCC is weighted by a score assigned to a corresponding category.

$$P_i(1\,|\,\theta_j) = \sum_k^{m+1} w_{ik} P_{ik}(1\,|\,\theta_j)$$

For more information about item calibration and determination, the reader is referred to Lord and Novick (1968) or Hambleton and Swaminathan (1985).

5.2 EQUATING

Although 2005 was the first year for the NECAP, it was still necessary to perform an equating because multiple forms were administered. The purpose of equating across forms was to place all items on a common scale. This within-year equating was accomplished by using the anchor-test nonequivalent group design. In this design, no assumption is made regarding the equivalence of the student groups used in equating the test forms; rather, they were assumed to be naturally occurring groups. Anchor items (in this particular case common items) are used to achieve the comparability of the groups. For NECAP, this within-year equating was performed by way of concurrent calibration. That is, all student response records were combined into a single sparse data matrix for each grade/content combination. PARSCALE was used to calibrate the datasets according to the models specified earlier, and item parameter estimates (along with standard errors) were determined.

The only exception was writing where the test forms were pre-equated from the field test administration (see 5.3 Standard Setting). However, the same IRT models were used for writing as were used in all other grade/contents (i.e., 3PL and GRM). Additionally, the same equating method was used (concurrent calibration) to put all test forms onto a common scale.

The next administration of NECAP will be scaled to the 2005 administration by way of equating. The particular equating procedure has yet to be determined and will be evaluated by the NECAP Technical Advisory Committee. It is reasonable that several different equating methods be compared for stability and for violations to any statistical assumptions underlying the equating methods.

The test designs displayed in Chapter 2 include a set of equating items that mirror the common test in terms of item types and distribution of emphasis. The set of equating items is matrixed across the forms of the test.

PRE-EQUATING FOR WRITING

Our psychometricians and test development staff worked closely and in an iterative fashion to develop parallel tests forms for writing. This process was a delicate balancing act between optimizing content and the psychometric properties of the test forms. As a result of this work, a report was generated and is included as Appendix C. By pre-equating the writing forms, a look-up table can be constructed in advance of any test administration. That is, as we have already estimated all item parameters, we can build the TCCs and the corresponding scaled scores for each raw score value. After the test is administered and a raw score is determined for a student, the look-up table is used to determine scaled scores and achievement levels.

5.3 STANDARD SETTING

A standard setting meeting was conducted for NECAP in January 2006. Thus, operational 2005 data were used to set standards for this assessment program, and 2005 will serve as the base year. All subsequent administrations of NECAP will be equated back to the 2005 scale.

The standard-setting report is included as Appendix D of this document. This detailed report outlines the methods and results of the standard-setting meetings. The results from the standard-setting meetings were the cut scores on the θ metric. Because the equating will scale back to the 2005 θ metric, the cut scores (Table 5-2) will remain fixed throughout the assessment program (unless standards are reset for any reason). After the standard-setting meetings were completed and the cut scores were determined, a meeting was held for the commissioners of education from each of the three states to review and officially adopt the final cutscores.

A list of Standard-Setting Committee member names and affiliations are included in Appendix A.

5.4 REPORTED SCALE SCORES

DESCRIPTION OF SCALE

A scale was developed for reporting purposes for each NECAP examination. These reporting scales are simple linear transformations of the underlying scale (θ) used in the IRT calibrations. The scales were developed such that for grade 3 the scaled scores ranged from 300 to 380, for grade 4 the range was 400 through 480, and so forth through grade 8, where scores ranged from 800 through 880. Likewise, the lowest scaled score in the *Proficient* range was placed at "X40" for each grade level. In other words, to be classified in the *Proficient* achievement level or above, a scaled score of 340 was required at grade 3, 440 at grade 4, and so forth.

Scaled scores supplement the achievement-level results by providing information about the position of a student's results within an achievement level. School- and district-level scaled scores are calculated by computing the average of student-level scaled scores. Students' raw scores, or total number of points, on the NECAP tests are translated to scaled scores using a data analysis process called *scaling*. Scaling simply converts raw points from one scale to another through the test characteristic curve. In the same way that the same temperature can be expressed on either the Fahrenheit or Celsius scales and the same distance can be expressed in either miles or kilometers, student scores on the NECAP tests could be expressed as raw scores (i.e., number correct) or scaled scores.

It is important to note that converting from raw scores to scaled scores does not change students' achievement-level classifications. Given the relative simplicity of raw scores, it is fair to question why scaled scores reports are used instead of raw scores in NECAP. Foremost, scaled scores offer the advantage of simplifying the reporting of results across content areas and subsequent years. Because the standard-setting process typically results in different cut scores across content areas on a raw score basis, it is useful to transform these raw cut scores to a scale that is more easily interpretable and consistent. For the NECAP, a score of, say, 340 for grade 3 is just beyond the range of scores associated with the *Partially Proficient* achievement level (placing the student in *Proficient*). This is true regardless of the content area or year with which one may be concerned. If one were to use raw scores, the raw cut score between *Partially Proficient* and *Proficient* may be, for example, 35 in mathematics but 33 in reading. Using scaled scores greatly simplifies the task of understanding how a student performed. Additionally, as scaled scores are linear transformations of θ and as the θ scale is where the equating is performed, scaled scores are comparable from one year to the next. The same cannot be said for simple raw scores.

CALCULATIONS

The scaled scores are obtained by a simple translation of ability estimates ($\hat{\theta}$) using the linear relationship between threshold values on the θ metric and their equivalent values on the scaled score metric. Students' ability estimates are based on their raw scores and are found by mapping through the TCC (as discussed). Scaled scores are calculated using the linear equation

$$SS = m\hat{\theta} + b$$

where *m* is the slope and *b* is the intercept. A separate linear transformation was used for each grade/content combination. Each line was determined by fixing the X40 value and the bottom of the scale (e.g., 300 for grade 3) to a location on the θ scale that was beyond the scaling of all the items across the various grade/content combinations. To determine this location, chance level (approximately equal to a student's expected performance by guessing) was compared to a value of –4.0 on the θ scale. Results of the examination were similar across all content areas; thus, for ease of communication and programming, the bottom of each scale was fixed at a θ = –4.0. Additionally, a raw score of 0 was assigned a scaled score of 300, and the maximum raw score was assigned a scaled score of 380 (both in the case of grade 3, whereas other grades used the same convention).

Because only two points within the θ scaled-score space were fixed, the cut scores between *Substantially Below Proficient* (SBP) and *Partially Proficient* (PP) and between *Proficient* (P) and *Proficient with Distinction* (PWD) varied across the grade/content combinations. Table 5-1 represents the scaled scores for each grade/content combination (i.e., the minimum scaled score for getting into the next achievement level). It is important to note that the values in Table 5-1 will not change from year to year because the cut scores along the θ scale will not change. In any given year, it may not be possible to attain a particular scaled score, but the scaled score cuts will remain the same.

Table 5-1
Scaled Scores for Each Achievement
Level Across All Grade/Content Combinations

Grade	Content	Min	Scale Score Cuts			Max
			SBP/PP	*PP/P*	*P/PWD*	
3	MAT	300	332	340	353	380
4	MAT	400	431	440	455	480
5	MAT	500	533	540	554	580
6	MAT	600	633	640	653	680
7	MAT	700	734	740	752	780
8	MAT	800	834	840	852	880
3	REA	300	331	340	357	380
4	REA	400	431	440	456	480
5	REA	500	530	540	556	580
6	REA	600	629	640	659	680
7	REA	700	729	740	760	780
8	REA	800	828	840	859	880
5	WRI	500	528	540	555	580
8	WRI	800	829	840	857	880

Table 5-2 shows the cut scores on the θ metric resulting from standard setting and the slope and intercept terms used to calculate the scaled scores. Here it is important to note that no number in Table 5-2 will change unless the standards are reset.

Appendix E contains the raw score to scaled score conversion tables. These are the actual tables that were used to determine student scaled scores (along with error bands) and achievement levels.

Table 5-2
Cutscores on θ Metric
Along with Slope and Intercept Terms

		θ Cuts				
Grade	Content	*SBP/PP*	*PP/P*	*P/PWD*	Intercept	Slope
3	MAT	–1.0381	–0.2685	0.9704	342.8782	10.7195
4	MAT	–1.1504	–0.37785	0.9493	444.1727	11.0432
5	MAT	–0.9279	–0.28455	1.0313	543.0634	10.7659
6	MAT	–0.87425	–0.22365	1.03425	642.3690	10.5922
7	MAT	–0.70795	–0.0787	1.09945	740.8028	10.2007
8	MAT	–0.6444	–0.0286	1.11775	840.2881	10.0720
3	REA	–1.32285	–0.497	1.0307	345.6751	11.4188
4	REA	–1.173	–0.3142	1.14725	443.4098	10.8525
5	REA	–1.33545	–0.4276	1.04035	544.7878	11.1970
6	REA	–1.47795	–0.51795	1.12545	645.9499	11.4875
7	REA	–1.4833	–0.5223	1.20575	746.0074	11.5019
8	REA	–1.52505	–0.5224	1.1344	846.0087	11.5022
5	WRI	–1.2008	–0.0232	1.5163	540.2334	10.0583
8	WRI	–1.0674	–0.0914	1.823	839.1064	9.7766

DISTRIBUTIONS

Appendix F of this document contains the scaled score cumulative density functions. These distributions were calculated using the sparse data matrix files that were used in the IRT calibrations. For each grade/content, these distributions show the cumulative percentage of students across the entire scaled score range.

SECTION II: STATISTICAL AND PSYCHOMETRIC SUMMARIES

CHAPTER 6—ITEM ANALYSES

As noted in Brown (1983), "A test is only as good as the items it contains." A complete evaluation of a test's quality must include an evaluation of each question. Both the *Standards for Educational and Psychological Testing* (AERA, 1999) and the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988) include standards for identifying quality questions. Questions should assess only knowledge or skills that are identified as part of the domain being measured and should avoid assessing irrelevant factors. They should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. Further, questions must not unfairly disadvantage test takers from particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses are conducted to ensure that NECAP questions meet these standards. Qualitative analyses were discussed in Chapter 2–Development and Test Design. The current section focuses on three categories of the quantitative evaluations: (1) difficulty indices, (2) item-test correlations, and (3) subgroup differences in item performance. The statistics presented in this section are based on 2005 NECAP results.

6.1 DIFFICULTY INDICES

All items were evaluated in terms of difficulty according to standard classical test theory practice. The expected item difficulty, also known as the *p-value*, is the main index of item difficulty under the classical test theory framework. This index measures difficulty by averaging the proportion of points received across all students who received the item. MC items were scored dichotomously (correct vs. incorrect), so for these items the difficulty index is simply the proportion of students who correctly answered the item. To place all item types on the same 0–1 scale, the p-value of an OR item was computed as the average score on the item divided by its maximum possible score. Although the p-value is traditionally described as a measure of difficulty (as it is described here), it is properly interpreted as an easiness index because larger values indicate easier items. An index of 0 indicates that no student received credit for the item, and an index of 1 indicates that every student received full credit for the item.

Items that are answered correctly by almost all students provide little information about differences in student ability, but they do indicate knowledge or skills that have been mastered by most students. Similarly, items that are correctly answered by very few students provide little information about differences in student ability but may indicate knowledge or skills that have not yet been mastered by most students. In general, to provide the most precise measurement, difficulty indices should range from near-chance performance (0.25 for four-option, MC items or essentially 0 for CR items) to 0.90. Experience has indicated that items conforming to this guideline tend to provide satisfactory statistical information for the bulk of the student population. However, on a criterion-referenced test, such as NECAP, it may be appropriate to include some items with difficulty values outside this region in order to measure well throughout the range of skill present at a given grade. Having a range of item

difficulties also helps to ensure that the test does not exhibit an excess of scores at the floor or ceiling of the distribution.

6.2 ITEM–TEST CORRELATIONS

A desirable feature of an item is that higher-ability students perform better on the item than do lower-ability students. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of an item. Within classical test theory, the item-test correlation is referred to as the *item's discrimination* because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. For polytomous items, the item discrimination index used was the Pearson product-moment correlation; for dichotomous items, the corresponding statistic is commonly called a *point-biserial correlation*. The theoretical range of these statistics is –1 to +1, with a typical range from 0.2 to 0.6.

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score; that is, the discrimination index can be interpreted as a measure of construct consistency. In light of this interpretation, the selection of an appropriate criterion total score is important to the interpretation of the discrimination index. For the 2005 NECAP, the item-test correlation was computed for each common item results are summarized in Chapter 6, subsection 6.2. The criterion score for these items was the student raw score (i.e., the sum of scores on the common items).

6.3 SUMMARY OF ITEM ANALYSIS RESULTS

Summary statistics of the difficulty and discrimination indices by grade and content area are provided in Appendix G as Tables G-1, G-2, and G-3. Table G-1 gives summary statistics for item difficulties and discriminations by form. In particular, Table G-1 displays the number of items, mean and standard deviation of the p-values, and mean and standard deviation of the discriminations by form. Table G-2 summarizes the p-values and item-total correlations by item type; specifically, the means and standard deviations of these statistics are provided for each item type (MC and OR) and are aggregated over both item types (ALL). Table G-2 also displays the number of items of each type (N). Note that for the p-values and discriminations in Table G-2, the numbers inside parentheses represent standard deviations and the numbers outside represent means. Finally, Table G-3 shows the number of common items that have difficulty or discrimination values within a stated range. The percentage of items in the range and cumulative percentage are also given.

The statistics in Table G-1 show that for each grade, content area, and form of the 2005 NECAP administration, the mean p-value fell between 0.40 and 0.78, whereas standard deviations ranged from 0.11 to 0.24. Mean item-total discriminations ranged from 0.36 to 0.52, with their standard deviations taking on values between 0.06 and 0.17. Table G-2 indicates that mean p-values for MC items fell between 0.54 and 0.85 among the various grades and content areas. For OR items, mean p-values were between 0.34 and 0.72, and aggregating all items together, they were between 0.46 and 0.78. Again comparing the different grades and content areas, MC items had mean item-total correlations between 0.31 and 0.45; for OR items, mean item-total correlations were between 0.47 and 0.65; aggregating among all items, they were between 0.38 and 0.48. Finally, observing Table G-3, the item difficulty and discrimination indices were generally in expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Additionally, the positive discrimination indices indicate that students

who performed well on individual items tended to perform well overall. Though it is not inappropriate to include items with low discrimination values or with very high or very low item difficulty values to ensure that the entire ability spectrum is appropriately covered, there were very few such cases on the NECAP assessments.

A comparison of indices across grade levels is complicated because these indices are population-dependent. Direct comparisons would require that either the items or students were common across groups. As that was not the case, it cannot be determined whether differences in performance across grade levels were due to differences in student ability or differences in item difficulty or both. However, one noteworthy statistical trend in math was that p-values tended to be highest at the lower grades.

Comparing the difficulty indices of MC and OR items is inappropriate because MC items can be answered correctly by guessing. Thus, it is not surprising that the p-values for MC items were higher than the p-values for OR items. Similarly, the partial credit allowed for OR items is advantageous in the computation of item-test correlations, so the discrimination indices for these items tended to be larger than the discrimination indices of MC items.

6.4 DIFFERENTIAL ITEM FUNCTIONING

The *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988) explicitly states that subgroup differences in performance should be examined when sample sizes permit and actions should be taken to make certain that differences in performance are due to construct-relevant, rather than irrelevant, factors. The *Standards for Educational and Psychological Testing* (AERA, 1999) includes similar guidelines. As part of the effort to identify such problems, NECAP items were evaluated in terms of DIF statistics.

DIF procedures are designed to identify items for which subgroups of interest perform differently beyond the impact of differences in overall achievement. For NECAP, the standardization DIF procedure (Dorans & Kulick, 1986) was employed to evaluate subgroup differences. This procedure calculates the difference in item performance for groups of students matched for achievement on the total test. That is, the average item performance is calculated for students at every total score; then an overall average is calculated weighting the total score distribution so it is the same for the two groups. In the 2005 NECAP DIF analysis, the criterion (matching) score for common items was the sum of scores on common items; the criterion score for matrix items was the sum of scores on common and matrix items (not including field-test items). Based on experience, this dual definition of the criterion score has worked well in identifying problematic common and matrix items.

The DIF index ranges from –1 to 1 for MC items and is adjusted to the same scale for OR items. Dorans and Holland (1993) suggested that index values between –0.05 and 0.05 should be considered negligible. Most NECAP items fell within this range, which is denoted "Type A" DIF in the tables presented in this section. Dorans and Holland further stated that items with values between –0.10 and –0.05 or between 0.05 and 0.10 (i.e., "low" or "Type B" DIF) should be inspected to ensure that no possible effect is overlooked and that items with values outside the (–0.10, 0.10) range (i.e., "high" or "Type C" DIF) are more unusual and should be examined very carefully.

DIF indices indicate differential performance between two groups. That differential performance may or may not be indicative of bias in the test. Course-taking patterns, group differences in interests, or differences in school curricula can lead to DIF. If subgroup differences in performance are related to construct-relevant factors, the items should be considered for inclusion on a test.

Each item was categorized according to the guidelines adapted from Dorans and Holland (1993). Tables 6-1, 6-2, and 6-3, which begin on page 71, present the number of items classified into each DIF

category separately by item type (MC or OR) and by overall (ALL). The three tables give results for male/female, white/black, and white/Hispanic comparisons, respectively, and are broken down by grade, content area, and form. Note that "Form 00" contains the common items that are used in calculating reported scores for students. The different DIF categories (Type A, B, or C) were defined above; a "Type D" designation indicates that there were not enough students in the grouping (fewer than 200 in at least one of the subgroups) to perform a reliable DIF analysis. In Table 6-2, blank values in the Type A, B, and C cells indicate that all items of the given form were "Type D" items. Table 6-4 presents the number of items classified into each DIF category by direction for the male/female comparison. For instance, the "F_A" column denotes the number of items that fell into the "A" range of DIF and for which females performed better than males, relative to performance on the test as a whole. Similarly, the "M_A" column gives the number of items that fell into the "A" range of DIF and for which males performed better than females, relative to performance on the test as a whole. The "N_A" column and "P_A" column display the number and proportion of items, respectively, in the "A" range. Results are disaggregated by grade, content area, and item type. To provide a complete summary across items, both common and matrix items are included in the tally of items falling into each category. The analysis in Table 6-4 was examined only for the male/female comparison because both the "male" and "female" subgroups exhibited high sample sizes; other comparisons had at least one subgroup with a substantially smaller sample size.

All of the tables presented in this section demonstrate that the majority of DIF distinctions in the 2005 NECAP examinations were in the "Type A" category and were thus characterized by "negligible" DIF levels as measured by Dorans and Holland (1993). Additional DIF analyses are located in Appendix H. Table H-1 shows the DIF categorization of each item by grade, content area, and form,

thus presenting a comprehensive list of DIF results for the 2005 NECAP examinations. Results are presented for the following comparisons:

- Male / female (M/F)
- White / black (W/B)
- White / Hispanic (W/H)
- White / Other (W/O)
- Rhode Island / Vermont (RI/VT)
- Rhode Island / New Hampshire (RI/NH)
- Vermont / New Hampshire (VT/NH)
- IEP / Non-IEP (IEP/NonIEP)
- LEP / Non-LEP (LEP/NonLEP)
- Low SES / High SES (SES/NonSES)

In Table H-1, the "Stat" columns give the value of the standardized DIF statistic (Dorans & Kulick, 1986), while the "Cat" columns indicate which DIF category each item fell into (Type A, B, or C) according to the criteria of Dorans and Holland (1993). Negative numbers in the "Stat" column indicate that the item was more difficult for the second subgroup listed in the column heading (e.g., female or non white students), relative to the test as a whole; positive numbers indicate that the item was relatively easier for these groups. Values are presented only when each subgroup in question was represented by at least 200 students; when the sample size of at least one subgroup did not reach this threshold, the appropriate cells were left blank. In some cases, the sparseness of the score distribution caused missing data at certain values of the criterion score; thus, the results of this section should be interpreted with caution.

Even though there are some items with DIF indices in the "low" or "high" categories, this does not necessarily indicate that the items are biased. Both the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988) and the *Standards for Educational and Psychological Testing* (AERA, 1999) assert that test items must be free from construct-irrelevant sources of differential difficulty. If subgroup differences in performance can be plausibly attributed to construct-relevant factors, the items may be included on a test. What is important is to determine whether the cause of this differential performance is construct-relevant

Table 6-1
Number of Items Classified into Each DIF Category, Male/Female

Gr.	Sub.	Form	All_A	All_B	All_C	All_D	MC_A	MC_B	MC_C	MC_D	OR_A	OR_B	OR_C	OR_D
03	MAT	00	52	3	0	0	33	2	0	0	19	1	0	0
03	MAT	01	8	2	0	0	4	2	0	0	4	0	0	0
03	MAT	02	9	0	0	0	6	0	0	0	3	0	0	0
03	MAT	03	9	1	0	0	5	1	0	0	4	0	0	0
03	MAT	04	8	1	0	0	5	1	0	0	3	0	0	0
03	MAT	05	11	0	0	0	6	0	0	0	5	0	0	0
03	MAT	06	7	2	0	0	4	2	0	0	3	0	0	0
03	MAT	07	9	1	0	0	5	1	0	0	4	0	0	0
03	MAT	08	9	0	0	0	6	0	0	0	3	0	0	0
03	MAT	09	9	1	0	0	5	1	0	0	4	0	0	0
03	REA	00	34	0	0	0	28	0	0	0	6	0	0	0
03	REA	01	17	0	0	0	14	0	0	0	3	0	0	0
03	REA	02	15	2	0	0	13	1	0	0	2	1	0	0
03	REA	03	16	1	0	0	13	1	0	0	3	0	0	0
04	MAT	00	47	7	1	0	30	4	1	0	17	3	0	0
04	MAT	01	10	0	0	0	6	0	0	0	4	0	0	0
04	MAT	02	10	0	0	0	6	0	0	0	4	0	0	0
04	MAT	03	9	1	0	0	5	1	0	0	4	0	0	0
04	MAT	04	9	1	0	0	6	0	0	0	3	1	0	0
04	MAT	05	8	1	0	0	5	1	0	0	3	0	0	0
04	MAT	06	9	1	0	0	5	1	0	0	4	0	0	0
04	MAT	07	9	1	0	0	5	1	0	0	4	0	0	0
04	MAT	08	10	0	0	0	6	0	0	0	4	0	0	0
04	MAT	09	9	1	0	0	5	1	0	0	4	0	0	0
04	REA	00	30	4	0	0	25	3	0	0	5	1	0	0
04	REA	01	15	2	0	0	12	2	0	0	3	0	0	0
04	REA	02	16	1	0	0	13	1	0	0	3	0	0	0
04	REA	03	15	2	0	0	13	1	0	0	2	1	0	0
05	MAT	00	41	7	0	0	25	7	0	0	16	0	0	0
05	MAT	01	11	0	0	0	6	0	0	0	5	0	0	0
05	MAT	02	9	2	0	0	6	0	0	0	3	2	0	0
05	MAT	03	9	2	0	0	5	1	0	0	4	1	0	0
05	MAT	04	11	0	0	0	6	0	0	0	5	0	0	0
05	MAT	05	10	1	0	0	5	1	0	0	5	0	0	0
05	MAT	06	10	1	0	0	6	0	0	0	4	1	0	0
05	MAT	07	11	0	0	0	6	0	0	0	5	0	0	0
05	MAT	08	11	0	0	0	6	0	0	0	5	0	0	0
05	MAT	09	9	2	0	0	5	1	0	0	4	1	0	0
05	REA	00	31	3	0	0	26	2	0	0	5	1	0	0
05	REA	01	13	4	0	0	10	4	0	0	3	0	0	0
05	REA	02	14	3	0	0	11	3	0	0	3	0	0	0
05	REA	03	16	1	0	0	13	1	0	0	3	0	0	0
05	WRI	01	17	0	0	0	10	0	0	0	7	0	0	0

Gr.	Sub.	Form	All_A	All_B	All_C	All_D	MC_A	MC_B	MC_C	MC_D	OR_A	OR_B	OR_C	OR_D
06	MAT	00	41	7	0	0	27	5	0	0	14	2	0	0
06	MAT	01	10	1	0	0	5	1	0	0	5	0	0	0
06	MAT	02	9	1	1	0	5	0	1	0	4	1	0	0
06	MAT	03	10	1	0	0	5	1	0	0	5	0	0	0
06	MAT	04	10	0	1	0	5	0	1	0	5	0	0	0
06	MAT	05	10	1	0	0	6	0	0	0	4	1	0	0
06	MAT	06	9	2	0	0	5	1	0	0	4	1	0	0
06	MAT	07	10	1	0	0	5	1	0	0	5	0	0	0
06	MAT	08	9	1	1	0	5	0	1	0	4	1	0	0
06	MAT	09	9	2	0	0	5	1	0	0	4	1	0	0
06	REA	00	30	4	0	0	25	3	0	0	5	1	0	0
06	REA	01	14	2	1	0	11	2	1	0	3	0	0	0
06	REA	02	13	3	1	0	10	3	1	0	3	0	0	0
06	REA	03	13	2	2	0	10	2	2	0	3	0	0	0
07	MAT	00	40	8	0	0	27	5	0	0	13	3	0	0
07	MAT	01	10	0	1	0	5	0	1	0	5	0	0	0
07	MAT	02	6	4	0	0	4	2	0	0	2	2	0	0
07	MAT	03	10	1	0	0	5	1	0	0	5	0	0	0
07	MAT	04	10	1	0	0	5	1	0	0	5	0	0	0
07	MAT	05	7	3	1	0	3	2	1	0	4	1	0	0
07	MAT	06	8	1	1	0	4	1	1	0	4	0	0	0
07	MAT	07	10	0	1	0	5	0	1	0	5	0	0	0
07	MAT	08	7	3	0	0	5	1	0	0	2	2	0	0
07	MAT	09	10	1	0	0	5	1	0	0	5	0	0	0
07	REA	00	23	9	2	0	19	7	2	0	4	2	0	0
07	REA	01	15	2	0	0	14	0	0	0	1	2	0	0
07	REA	02	14	3	0	0	11	3	0	0	3	0	0	0
07	REA	03	14	1	2	0	12	0	2	0	2	1	0	0
08	MAT	00	40	8	0	0	27	5	0	0	13	3	0	0
08	MAT	01	7	3	1	0	3	3	0	0	4	0	1	0
08	MAT	02	8	2	0	0	5	1	0	0	3	1	0	0
08	MAT	03	10	1	0	0	5	1	0	0	5	0	0	0
08	MAT	04	9	2	0	0	5	1	0	0	4	1	0	0
08	MAT	05	9	1	0	0	5	1	0	0	4	0	0	0
08	MAT	06	9	2	0	0	6	0	0	0	3	2	0	0
08	MAT	07	8	3	0	0	4	2	0	0	4	1	0	0
08	MAT	08	7	3	0	0	5	1	0	0	2	2	0	0
08	MAT	09	9	2	0	0	4	2	0	0	5	0	0	0
08	REA	00	24	6	4	0	21	3	4	0	3	3	0	0
08	REA	01	11	6	0	0	10	4	0	0	1	2	0	0
08	REA	02	11	3	3	0	10	1	3	0	1	2	0	0
08	REA	03	13	4	0	0	12	2	0	0	1	2	0	0
08	WRI	01	15	2	0	0	8	2	0	0	7	0	0	0

Table 6-2
Number of Items Classified into Each DIF Category, White/Black Comparison

Gr.	Sub.	Form	All_A	All_B	All_C	All_D	MC_A	MC_B	MC_C	MC_D	OR_A	OR_B	OR_C	OR_D
03	MAT	00	49	6	0	0	31	4	0	0	18	2	0	0
03	MAT	01				10				6				4
03	MAT	02				9				6				3
03	MAT	03				10				6				4
03	MAT	04				9				6				3
03	MAT	05				11				6				5
03	MAT	06				9				6				3
03	MAT	07				10				6				4
03	MAT	08				9				6				3
03	MAT	09				10				6				4
03	REA	00	31	3	0	0	26	2	0	0	5	1	0	0
03	REA	01				17				14				3
03	REA	02				17				14				3
03	REA	03				17				14				3
04	MAT	00	51	4	0	0	33	2	0	0	18	2	0	0
04	MAT	01				10				6				4
04	MAT	02				10				6				4
04	MAT	03				10				6				4
04	MAT	04				10				6				4
04	MAT	05				9				6				3
04	MAT	06				10				6				4
04	MAT	07				10				6				4
04	MAT	08				10				6				4
04	MAT	09				10				6				4
04	REA	00	28	6	0	0	22	6	0	0	6	0	0	0
04	REA	01				17				14				3
04	REA	02				17				14				3
04	REA	03				17				14				3
05	MAT	00	42	6	0	0	30	2	0	0	12	4	0	0
05	MAT	01				11				6				5
05	MAT	02				11				6				5
05	MAT	03				11				6				5
05	MAT	04				11				6				5
05	MAT	05				11				6				5
05	MAT	06				11				6				5
05	MAT	07				11				6				5
05	MAT	08				11				6				5
05	MAT	09				11				6				5
05	REA	00	26	8	0	0	20	8	0	0	6	0	0	0
05	REA	01				17				14				3
05	REA	02				17				14				3
05	REA	03				17				14				3
05	WRI	01	13	4	0	0	6	4	0	0	7	0	0	0
06	MAT	00	47	1	0	0	32	0	0	0	15	1	0	0

Gr.	Sub.	Form	All_A	All_B	All_C	All_D	MC_A	MC_B	MC_C	MC_D	OR_A	OR_B	OR_C	OR_D
06	MAT	01				11				6				5
06	MAT	02				11				6				5
06	MAT	03				11				6				5
06	MAT	04				11				6				5
06	MAT	05				11				6				5
06	MAT	06				11				6				5
06	MAT	07				11				6				5
06	MAT	08				11				6				5
06	MAT	09				11				6				5
06	REA	00	32	2	0	0	26	2	0	0	6	0	0	0
06	REA	01				17				14				3
06	REA	02				17				14				3
06	REA	03				17				14				3
07	MAT	00	41	7	0	0	28	4	0	0	13	3	0	0
07	MAT	01				11				6				5
07	MAT	02				10				6				4
07	MAT	03				11				6				5
07	MAT	04				11				6				5
07	MAT	05				11				6				5
07	MAT	06				10				6				4
07	MAT	07				11				6				5
07	MAT	08				10				6				4
07	MAT	09				11				6				5
07	REA	00	28	6	0	0	22	6	0	0	6	0	0	0
07	REA	01				17				14				3
07	REA	02				17				14				3
07	REA	03				17				14				3
08	MAT	00	44	4	0	0	30	2	0	0	14	2	0	0
08	MAT	01				11				6				5
08	MAT	02				10				6				4
08	MAT	03				11				6				5
08	MAT	04				11				6				5
08	MAT	05				10				6				4
08	MAT	06				11				6				5
08	MAT	07				11				6				5
08	MAT	08				10				6				4
08	MAT	09				11				6				5
08	REA	00	28	3	3	0	22	3	3	0	6	0	0	0
08	REA	01				17				14				3
08	REA	02				17				14				3
08	REA	03				17				14				3
08	WRI	01	13	4	0	0	7	3	0	0	6	1	0	0

Table 6-3
Number of Items Classified into Each DIF Category by Form, White/Hispanic Comparison

Gr.	Sub.	Form	All_A	All_B	All_C	All_D	MC_A	MC_B	MC_C	MC_D	OR_A	OR_B	OR_C	OR_D
03	MAT	00	38	17	0	0	25	10	0	0	13	7	0	0
03	MAT	01	7	3	0	0	4	2	0	0	3	1	0	0
03	MAT	02	4	5	0	0	3	3	0	0	1	2	0	0
03	MAT	03	9	1	0	0	6	0	0	0	3	1	0	0
03	MAT	04	6	2	1	0	5	1	0	0	1	1	1	0
03	MAT	05	6	4	1	0	4	1	1	0	2	3	0	0
03	MAT	06	6	2	1	0	5	0	1	0	1	2	0	0
03	MAT	07	8	1	1	0	4	1	1	0	4	0	0	0
03	MAT	08	6	2	1	0	4	1	1	0	2	1	0	0
03	MAT	09	8	2	0	0	4	2	0	0	4	0	0	0
03	REA	00	24	9	1	0	18	9	1	0	6	0	0	0
03	REA	01	14	1	2	0	11	1	2	0	3	0	0	0
03	REA	02	13	3	1	0	10	3	1	0	3	0	0	0
03	REA	03	15	1	1	0	12	1	1	0	3	0	0	0
04	MAT	00	44	9	2	0	32	2	1	0	12	7	1	0
04	MAT	01	5	3	2	0	3	3	0	0	2	0	2	0
04	MAT	02	7	3	0	0	4	2	0	0	3	1	0	0
04	MAT	03	7	1	2	0	5	0	1	0	2	1	1	0
04	MAT	04	7	3	0	0	5	1	0	0	2	2	0	0
04	MAT	05	4	4	1	0	3	2	1	0	1	2	0	0
04	MAT	06	6	4	0	0	2	4	0	0	4	0	0	0
04	MAT	07	4	6	0	0	3	3	0	0	1	3	0	0
04	MAT	08	4	5	1	0	3	2	1	0	1	3	0	0
04	MAT	09	7	2	1	0	3	2	1	0	4	0	0	0
04	REA	00	24	6	4	0	19	5	4	0	5	1	0	0
04	REA	01	14	3	0	0	11	3	0	0	3	0	0	0
04	REA	02	12	4	1	0	9	4	1	0	3	0	0	0
04	REA	03	8	8	1	0	6	7	1	0	2	1	0	0
05	MAT	00	41	7	0	0	28	4	0	0	13	3	0	0
05	MAT	01	7	4	0	0	3	3	0	0	4	1	0	0
05	MAT	02	9	2	0	0	4	2	0	0	5	0	0	0
05	MAT	03	6	4	1	0	2	3	1	0	4	1	0	0
05	MAT	04	10	1	0	0	5	1	0	0	5	0	0	0
05	MAT	05	6	5	0	0	4	2	0	0	2	3	0	0
05	MAT	06	8	1	2	0	6	0	0	0	2	1	2	0
05	MAT	07	10	1	0	0	5	1	0	0	5	0	0	0
05	MAT	08	9	2	0	0	6	0	0	0	3	2	0	0
05	MAT	09	8	2	1	0	4	1	1	0	4	1	0	0
05	REA	00	22	9	3	0	16	9	3	0	6	0	0	0
05	REA	01	10	7	0	0	8	6	0	0	2	1	0	0
05	REA	02	11	4	2	0	9	3	2	0	2	1	0	0
05	REA	03	8	6	3	0	5	6	3	0	3	0	0	0
05	WRI	01	13	3	1	0	6	3	1	0	7	0	0	0

Gr.	Sub.	Form	All_A	All_B	All_C	All_D	MC_A	MC_B	MC_C	MC_D	OR_A	OR_B	OR_C	OR_D
06	MAT	00	43	4	1	0	29	3	0	0	14	1	1	0
06	MAT	01	8	2	1	0	4	1	1	0	4	1	0	0
06	MAT	02	8	2	1	0	5	1	0	0	3	1	1	0
06	MAT	03	8	3	0	0	5	1	0	0	3	2	0	0
06	MAT	04	8	3	0	0	6	0	0	0	2	3	0	0
06	MAT	05	9	1	1	0	4	1	1	0	5	0	0	0
06	MAT	06	7	3	1	0	4	1	1	0	3	2	0	0
06	MAT	07	9	2	0	0	5	1	0	0	4	1	0	0
06	MAT	08	6	4	1	0	3	3	0	0	3	1	1	0
06	MAT	09	8	3	0	0	6	0	0	0	2	3	0	0
06	REA	00	28	5	1	0	22	5	1	0	6	0	0	0
06	REA	01	14	2	1	0	11	2	1	0	3	0	0	0
06	REA	02	15	1	1	0	12	1	1	0	3	0	0	0
06	REA	03	13	2	2	0	10	2	2	0	3	0	0	0
07	MAT	00	41	7	0	0	26	6	0	0	15	1	0	0
07	MAT	01	10	1	0	0	6	0	0	0	4	1	0	0
07	MAT	02	8	2	0	0	6	0	0	0	2	2	0	0
07	MAT	03	7	3	1	0	5	1	0	0	2	2	1	0
07	MAT	04	8	3	0	0	3	3	0	0	5	0	0	0
07	MAT	05	9	2	0	0	5	1	0	0	4	1	0	0
07	MAT	06	9	0	1	0	5	0	1	0	4	0	0	0
07	MAT	07	8	3	0	0	4	2	0	0	4	1	0	0
07	MAT	08	8	2	0	0	4	2	0	0	4	0	0	0
07	MAT	09	8	3	0	0	4	2	0	0	4	1	0	0
07	REA	00	27	6	1	0	21	6	1	0	6	0	0	0
07	REA	01	8	5	4	0	7	3	4	0	1	2	0	0
07	REA	02	12	4	1	0	9	4	1	0	3	0	0	0
07	REA	03	15	2	0	0	12	2	0	0	3	0	0	0
08	MAT	00	45	3	0	0	30	2	0	0	15	1	0	0
08	MAT	01	8	2	1	0	5	1	0	0	3	1	1	0
08	MAT	02	9	1	0	0	6	0	0	0	3	1	0	0
08	MAT	03	8	3	0	0	4	2	0	0	4	1	0	0
08	MAT	04	9	2	0	0	4	2	0	0	5	0	0	0
08	MAT	05	9	1	0	0	5	1	0	0	4	0	0	0
08	MAT	06	8	3	0	0	6	0	0	0	2	3	0	0
08	MAT	07	6	4	1	0	4	2	0	0	2	2	1	0
08	MAT	08	7	3	0	0	3	3	0	0	4	0	0	0
08	MAT	09	7	3	1	0	2	3	1	0	5	0	0	0
08	REA	00	20	9	5	0	14	9	5	0	6	0	0	0
08	REA	01	12	4	1	0	10	3	1	0	2	1	0	0
08	REA	02	7	7	3	0	7	4	3	0	0	3	0	0
08	REA	03	15	2	0	0	12	2	0	0	3	0	0	0
08	WRI	01	12	3	2	0	6	2	2	0	6	1	0	0

Legend for Tables 6-1, 6-2 and 6-3

Gr. = Grade

Sub. = Subject

Form = Form Number

All_A = Number of items categorized by "Type A" DIF; includes both MC and OR items

All_B = Number of items categorized by "Type B" DIF; includes both MC and OR items

All_C = Number of items categorized by "Type C" DIF; includes both MC and OR items

All_D = Number of items categorized by "Type D" DIF (not enough students to perform a reliable DIF analysis); includes both MC and OR items

MC_A = Number of items categorized by "Type A" DIF; includes MC items only

MC_B = Number of items categorized by "Type B" DIF; includes MC items only

MC_C = Number of items categorized by "Type C" DIF; includes MC items only

MC_D = Number of items categorized by "Type D" DIF (not enough students to perform a reliable DIF analysis); includes MC items only

OR_A = Number of items categorized by "Type A" DIF; includes OR items only

OR_B = Number of items categorized by "Type B" DIF; includes OR items only

OR_C = Number of items categorized by "Type C" DIF; includes OR items only

OR_D = Number of items categorized by "Type D" DIF (not enough students to perform a reliable DIF analysis); includes OR items only

Table 6-4
Number and Proportion of Items Classified into Each DIF Category and Direction by Item Type, Male/Female Comparison

Grade	Subject	Type	F_A	M_A	N_A	P_A	F_B	M_B	N_B	P_B	F_C	M_C	N_C	P_C	N_D	P_D
03	MAT	MC	48	31	79	0.89	3	7	10	0.11	0	0	0	0	0	0
03	MAT	OR	27	25	52	0.98	1	0	1	0.02	0	0	0	0	0	0
03	REA	MC	47	21	68	0.97	1	1	2	0.03	0	0	0	0	0	0
03	REA	OR	10	4	14	0.93	1	0	1	0.07	0	0	0	0	0	0
04	MAT	MC	53	26	79	0.89	1	8	9	0.1	0	1	1	0.01	0	0
04	MAT	OR	34	17	51	0.93	3	1	4	0.07	0	0	0	0	0	0
04	REA	MC	32	31	63	0.9	0	7	7	0.1	0	0	0	0	0	0
04	REA	OR	8	5	13	0.87	2	0	2	0.13	0	0	0	0	0	0
05	MAT	MC	31	45	76	0.88	1	9	10	0.12	0	0	0	0	0	0
05	MAT	OR	39	17	56	0.92	1	4	5	0.08	0	0	0	0	0	0
05	REA	MC	18	42	60	0.86	1	9	10	0.14	0	0	0	0	0	0
05	REA	OR	13	1	14	0.93	1	0	1	0.07	0	0	0	0	0	0
05	WRI	MC	4	6	10	1	0	0	0	0	0	0	0	0	0	0
05	WRI	OR	7	0	7	1	0	0	0	0	0	0	0	0	0	0
06	MAT	MC	33	40	73	0.85	1	9	10	0.12	0	3	3	0.03	0	0
06	MAT	OR	24	30	54	0.89	4	3	7	0.11	0	0	0	0	0	0
06	REA	MC	19	37	56	0.8	0	10	10	0.14	0	4	4	0.06	0	0
06	REA	OR	12	2	14	0.93	1	0	1	0.07	0	0	0	0	0	0
07	MAT	MC	35	33	68	0.79	1	13	14	0.16	0	4	4	0.05	0	0
07	MAT	OR	29	21	50	0.86	5	3	8	0.14	0	0	0	0	0	0
07	REA	MC	29	27	56	0.8	0	10	10	0.14	0	4	4	0.06	0	0
07	REA	OR	10	0	10	0.67	5	0	5	0.33	0	0	0	0	0	0
08	MAT	MC	38	31	69	0.8	1	16	17	0.2	0	0	0	0	0	0
08	MAT	OR	32	15	47	0.81	8	2	10	0.17	1	0	1	0.02	0	0
08	REA	MC	24	29	53	0.76	0	10	10	0.14	0	7	7	0.1	0	0
08	REA	OR	6	0	6	0.4	9	0	9	0.6	0	0	0	0	0	0
08	WRI	MC	6	2	8	0.8	0	2	2	0.2	0	0	0	0	0	0
08	WRI	OR	7	0	7	1	0	0	0	0	0	0	0	0	0	0

6.5 ITEM RESPONSE THEORY ANALYSES

Chapter 5, subsection 5.1 introduced the notion of IRT and gave a thorough description of the topic. As noted in that subsection, all NECAP items were calibrated using IRT, and the calibrated item parameters were ultimately used to scale both the items and students onto a common framework. The results of the analyses are presented in this subsection and its appendix, Appendix I.

Tables I-1 to I-28 in Appendix I give the IRT item parameters of all common items on the 2005 NECAP examinations, broken down by grade and content area. Additionally, the Test Characteristic Curve (TCC) and TIF are provided for each grade and content area. These are presented in Appendix I as Figures I-1 to I-28. The TCC displays the expected (average) raw score associated with each θ_j value between –4 and 4. Mathematically, the TCC is computed by summing the ICCs of all items that contribute to the raw score. Using the notation introduced in subsection 5.1, the expected raw score at a given value of θ_j is

$$E(X \mid \theta_j) = \sum_{i=1}^{n} P_i\left(1 \middle| \theta_j\right),$$

where i indexes the items (and n is the number of items contributing to the raw score),

j indexes students (here, θ_j runs from –4 to 4), and

$E(X \mid \theta_j)$ is the expected raw score for a student of ability θ_j.

The expected raw score is monotonic in that it increases with θ_j, consistent with the notion that students of high ability tend to earn higher raw scores than students of low ability. Most TCCs are "S-shaped" in that they are flatter at the ends of the distribution and steeper in the middle.

The TIF displays the amount of statistical information that the test provides at each value of θ_j. The statistical importance of TIFs is due to the fact that there is a direct relation between the information of a test and its standard error of measurement (SEM). Moreover, information functions are essentially

functions that depict test precision across the entire latent trait continuum. For long tests, the SEM at a given θ_j is approximately equal to the inverse of the square root of the statistical information at θ_j (Hambleton, Swaminathan, & Rogers, 1991):

$$SEM(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}}$$

Compared to the tails, TIFs are often higher near the middle of the θ distribution, where most students are located and hence where most items are designed to measure.

CHAPTER 7—RELIABILITY

Although an individual item's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way in which items function together and complement one another. Any measurement includes some amount of measurement error; that is, no measurement can be perfectly accurate. This is true of academic assessments: No assessment can measure student performance with perfect accuracy; some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. Items that function well together produce assessments that have less measurement error (that is, the errors made should be small on average). Such assessments are described as reliable.

There are a number of ways to estimate an assessment's reliability. One approach is to split all test items into two groups and then correlate students' scores on the two half-tests. This is known as a *split-half estimate of reliability*. If the two half-test scores correlate highly, items on the two half-tests are likely to be measuring very similar knowledge or skills. Such a correlation is evidence that the items complement one another and function well as a group, suggesting that measurement error will be minimal.

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation. Cronbach (1951) provided a statistic that avoids this concern about the split-half method. Cronbach's α coefficient is an estimate of the average of all possible split-half reliability coefficients. This statistic was used to assess the reliability of the 2005 NECAP examinations, as described in the following subsections.

7.1 RELIABILITY AND STANDARD ERRORS OF MEASUREMENT

Table 7-1 presents descriptive statistics, Cronbach's α coefficient, and raw score SEMs for each content area and grade (statistics are based on common items only). Cronbach's α is computed using the following formula:

$$\alpha \equiv \frac{n}{n-1}\left[1-\frac{\sum_{i=1}^{n}\sigma^2\left(Y_i\right)}{\sigma_x^2}\right]$$

where *i* indexes the item,

n is the number of items,

$\sigma^2\left(Y_i\right)$ represents individual item variance, and

σ_x^2 represents the total test variance.

Table 7-1
Reliabilities, Standard Errors of Measurement, and Descriptive Statistics

Grade	Subject	N	Points	Min	Max	Mean	S.D.	Rel. (α)	S.E.M.
03	MAT	32219	65	0	65	40.341	13.693	0.934	3.528
03	REA	32087	52	0	52	31.446	10.869	0.895	3.522
04	MAT	32673	65	0	65	39.628	13.043	0.925	3.574
04	REA	32527	52	0	52	33.452	9.112	0.891	3.01
05	MAT	33532	66	0	66	31.546	13.68	0.917	3.934
05	REA	33402	52	0	52	29.153	8.64	0.876	3.042
05	WRI	33376	37	0	37	22.483	5.168	0.72	2.734
06	MAT	34826	66	0	66	32.416	14.283	0.925	3.92
06	REA	34684	52	0	52	30.925	9.181	0.889	3.053
07	MAT	35336	66	0	65	30.596	12.507	0.903	3.89
07	REA	35245	52	0	52	30.507	9.197	0.891	3.037
08	MAT	36636	66	0	66	32.896	14.286	0.925	3.911
08	REA	36582	52	0	52	31.979	8.948	0.889	2.978
08	WRI	36444	37	0	37	23.04	5.97	0.73	3.102

For mathematics, the reliability coefficient ranged from 0.903 to 0.934; for reading, the coefficient ranged from 0.876 to 0.895; for the two writing examinations, the values were 0.72 and 0.73.

Because different grades and content areas have different test designs (e.g., the number of items varies by test), it is inappropriate to make inferences about the quality of one test by comparing its reliability to that of another test from a different grade and/or content area.

7.2 SUBGROUP RELIABILITY

In subsection 7.1, the reliability coefficients presented were calculated based on the overall population of students who took the NECAP assessment in 2005. Table 7-2 also presents Cronbach's α coefficient for various subgroups of interest. Each subgroup reliability was calculated via the same formula defined in subsection 7.1, using only the members of the subgroup in the computations.

Table 7-2
Reliabilities of Subgroups

Grade	Subject	Subgroup	N	Alpha
03	MAT	White	27257	0.93
03	MAT	Native Hawaiian or Pacific Islander	40	0.96
03	MAT	Hispanic or Latino	2447	0.92
03	MAT	Black or African American	1358	0.93
03	MAT	Asian	830	0.94
03	MAT	American Indian or Alaskan Native	129	0.93
03	MAT	LEP	1653	0.94
03	MAT	IEP	4168	0.93
03	MAT	Low SES	9458	0.93
03	REA	White	27220	0.89
03	REA	Native Hawaiian or Pacific Islander	40	0.94
03	REA	Hispanic or Latino	2391	0.88
03	REA	Black or African American	1342	0.89
03	REA	Asian	810	0.89
03	REA	American Indian or Alaskan Native	127	0.89
03	REA	LEP	1542	0.90
03	REA	IEP	4147	0.90
03	REA	Low SES	9387	0.89
04	MAT	White	27802	0.92
04	MAT	Native Hawaiian or Pacific Islander	51	0.94
04	MAT	Hispanic or Latino	2432	0.92
04	MAT	Black or African American	1351	0.93
04	MAT	Asian	783	0.93
04	MAT	American Indian or Alaskan Native	113	0.94
04	MAT	LEP	1583	0.93
04	MAT	IEP	4612	0.93
04	MAT	Low SES	9552	0.92
04	REA	White	27755	0.88
04	REA	Native Hawaiian or Pacific Islander	52	0.93
04	REA	Hispanic or Latino	2369	0.89
04	REA	Black or African American	1331	0.90
04	REA	Asian	769	0.89
04	REA	American Indian or Alaskan Native	113	0.90
04	REA	LEP	1463	0.89
04	REA	IEP	4575	0.89
04	REA	Low SES	9446	0.89
05	MAT	White	28710	0.91
05	MAT	Native Hawaiian or Pacific Islander	53	0.92
05	MAT	Hispanic or Latino	2391	0.89
05	MAT	Black or African American	1321	0.89
05	MAT	Asian	768	0.93
05	MAT	American Indian or Alaskan Native	133	0.90
05	MAT	LEP	1452	0.90
05	MAT	IEP	5096	0.89

Grade	Subject	Subgroup	N	Alpha
05	MAT	Low SES	9593	0.90
05	REA	White	28680	0.87
05	REA	Native Hawaiian or Pacific Islander	53	0.92
05	REA	Hispanic or Latino	2338	0.87
05	REA	Black or African American	1302	0.87
05	REA	Asian	744	0.87
05	REA	American Indian or Alaskan Native	132	0.89
05	REA	LEP	1338	0.87
05	REA	IEP	5053	0.86
05	REA	Low SES	9511	0.87
05	WRI	White	28675	0.71
05	WRI	Native Hawaiian or Pacific Islander	52	0.82
05	WRI	Hispanic or Latino	2329	0.78
05	WRI	Black or African American	1295	0.78
05	WRI	Asian	742	0.72
05	WRI	American Indian or Alaskan Native	131	0.70
05	WRI	LEP	1335	0.80
05	WRI	IEP	5066	0.77
05	WRI	Low SES	9493	0.75
06	MAT	White	29666	0.92
06	MAT	Native Hawaiian or Pacific Islander	70	0.94
06	MAT	Hispanic or Latino	2610	0.90
06	MAT	Black or African American	1374	0.91
06	MAT	Asian	774	0.94
06	MAT	American Indian or Alaskan Native	130	0.92
06	MAT	LEP	1320	0.91
06	MAT	IEP	5294	0.90
06	MAT	Low SES	9861	0.91
06	REA	White	29609	0.88
06	REA	Native Hawaiian or Pacific Islander	67	0.92
06	REA	Hispanic or Latino	2554	0.89
06	REA	Black or African American	1364	0.89
06	REA	Asian	758	0.90
06	REA	American Indian or Alaskan Native	130	0.90
06	REA	LEP	1217	0.88
06	REA	IEP	5247	0.87
06	REA	Low SES	9766	0.88
07	MAT	White	30209	0.90
07	MAT	Native Hawaiian or Pacific Islander	98	0.94
07	MAT	Hispanic or Latino	2541	0.87
07	MAT	Black or African American	1442	0.88
07	MAT	Asian	731	0.92
07	MAT	American Indian or Alaskan Native	145	0.88
07	MAT	LEP	1149	0.90
07	MAT	IEP	5335	0.86
07	MAT	Low SES	9561	0.88
07	REA	White	30208	0.88

Grade	Subject	Subgroup	N	Alpha
07	REA	Native Hawaiian or Pacific Islander	99	0.93
07	REA	Hispanic or Latino	2490	0.88
07	REA	Black or African American	1421	0.88
07	REA	Asian	712	0.90
07	REA	American Indian or Alaskan Native	144	0.90
07	REA	LEP	1024	0.89
07	REA	IEP	5336	0.88
07	REA	Low SES	9479	0.88
08	MAT	White	31780	0.92
08	MAT	Native Hawaiian or Pacific Islander	112	0.95
08	MAT	Hispanic or Latino	2421	0.91
08	MAT	Black or African American	1332	0.91
08	MAT	Asian	721	0.94
08	MAT	American Indian or Alaskan Native	148	0.91
08	MAT	LEP	978	0.91
08	MAT	IEP	5610	0.89
08	MAT	Low SES	9166	0.91
08	REA	White	31798	0.88
08	REA	Native Hawaiian or Pacific Islander	113	0.93
08	REA	Hispanic or Latino	2365	0.88
08	REA	Black or African American	1322	0.88
08	REA	Asian	712	0.90
08	REA	American Indian or Alaskan Native	149	0.87
08	REA	LEP	876	0.88
08	REA	IEP	5610	0.87
08	REA	Low SES	9107	0.88
08	WRI	White	31717	0.71
08	WRI	Native Hawaiian or Pacific Islander	107	0.83
08	WRI	Hispanic or Latino	2334	0.77
08	WRI	Black or African American	1306	0.75
08	WRI	Asian	713	0.74
08	WRI	American Indian or Alaskan Native	147	0.70
08	WRI	LEP	867	0.79
08	WRI	IEP	5563	0.75
08	WRI	Low SES	9042	0.74

For mathematics, subgroup reliabilities ranged from 0.86 to 0.96; for reading, they ranged from 0.86 to 0.94; for writing, they ranged from 0.70 to 0.83. The subgroup reliabilities for writing were lower than those of the other two content areas, but the two writing examinations (Grades 5 and 8) displayed a high degree of consistency with each other. Additionally, as stated in subsection 7.1, qualitative differences between grades and content areas preclude valid inference about the quality of a

test based on statistical comparison with other tests. In sum, no further investigation of any NECAP examination was warranted on the basis of the results presented in Table 7-2.

It is important to note that the reliabilities of different subgroups are also not comparable to one another due to the fact that their populations have varied statistical distributions. For example, a reliability coefficient, as a type of correlation coefficient, may be artificially low for subgroups whose populations have little variability (Draper & Smith, 1998). In other words, reliabilities are dependent not only on the measurement properties of a test but on the statistical distribution of the studied subgroup. Another reason why comparisons between subgroups are not advised is the fact that sample sizes of the different groups varied considerably (see the column of Table 7-2 entitled "N" for subgroup sample sizes). When computing the reliability coefficients of many subgroups that have markedly different sample sizes, some natural variation in the results is expected. Finally, there is no industry standard to interpret the strength of a reliability coefficient, and this is particularly true when the population of interest is a single subgroup. For all of these reasons, the results of this subsection should be interpreted with caution.

7.3 STRATIFIED COEFFICIENT ALPHA

According to Feldt and Brennan (1989), a prescribed distribution of items over categories (such as different item types) indicates the presumption that at least a small, but important, degree of unique variance is associated with the categories. In contrast, Cronbach's coefficient α is built on the assumption that there are no such local or clustered dependencies. A stratified version of coefficient α corrects for this problem:

$$\alpha_{strat} = 1 - \frac{\sum_{j=1}^{k} \sigma_{x_j}^2 (1-\alpha)}{\sigma_x^2}$$

where *j* indexes the subtests or categories,

$\sigma_{x_j}^2$ represents the variance of the *k* individual subtests or categories,

α is the unstratified Cronbach's α coefficient, and

σ_x^2 represents the total test variance.

Stratified coefficient α was calculated separately for each grade/content combination. The stratification was based on item types (MC v. OR). These results are provided in Table 7-3; all statistics in this table are based on common items only. From left to right, columns in Table 7-3 display the following information:

- Grade
- Subject
- Reliability coefficient (alpha) based on both MC and OR items
- Reliability coefficient based on MC items only
- The number of MC items
- Reliability coefficient based on OR items only
- The number of OR items (with the number of possible OR points in parentheses)
- Stratified alpha coefficient.

Table 7-3
Coefficients α and Stratified α_s by Grade and Content Area

Grade	Subject	Alpha	MC alpha	N (MC)	OR alpha	N (OR)	Strat Alpha
03	MAT	0.93	0.89	35	0.87	20 (30)	0.94
03	REA	0.9	0.89	28	0.79	6 (24)	0.91
04	MAT	0.92	0.88	35	0.85	20 (30)	0.93
04	REA	0.89	0.89	28	0.72	6 (24)	0.9
05	MAT	0.92	0.86	32	0.86	16 (34)	0.92
05	REA	0.88	0.84	28	0.78	6 (24)	0.89
05	WRI	0.72	0.7	10	0.65	7 (27)	0.75
06	MAT	0.92	0.86	32	0.87	16 (34)	0.93
06	REA	0.89	0.85	28	0.83	6 (24)	0.91
07	MAT	0.9	0.83	32	0.83	16 (34)	0.91
07	REA	0.89	0.85	28	0.86	6 (24)	0.91
08	MAT	0.93	0.87	32	0.87	16 (34)	0.93
08	REA	0.89	0.85	28	0.88	6 (24)	0.91
08	WRI	0.73	0.64	10	0.66	7 (27)	0.75

Table 7-4 presents the following statistics for each form:

- Reliability among both common and matrix items (Alpha)
- Reliability among MC items only (MC alpha)
- Reliability among OR items only (OR alpha)
- Reliability stratified by MC/OR (Frmt Strat)
- Reliability among common items only (Com alpha)

Note that if a test only has one form, then all cells between Form02 and Form09 are left blank.

Table 7-4
Reliability by Form and Item Type

Grade	Subject	Stat	Form01	Form02	Form03	Form04	Form05	Form06	Form07	Form08	Form09
03	MAT	Alpha	0.94	0.94	0.95	0.94	0.95	0.94	0.94	0.94	0.95
03	MAT	MC alpha	0.90	0.91	0.91	0.91	0.91	0.90	0.91	0.90	0.91
03	MAT	OR alpha	0.88	0.88	0.90	0.88	0.90	0.88	0.88	0.88	0.89
03	MAT	Frmt Strat	0.94	0.94	0.95	0.94	0.95	0.94	0.94	0.94	0.95
03	MAT	Com alpha	0.93	0.93	0.93	0.94	0.93	0.93	0.93	0.93	0.93
03	REA	Alpha	0.93	0.93	0.93	0.89	0.90	0.89	0.90	0.89	0.89
03	REA	MC alpha	0.92	0.92	0.92	0.89	0.89	0.88	0.89	0.88	0.88
03	REA	OR alpha	0.86	0.85	0.86	0.79	0.79	0.79	0.79	0.78	0.78
03	REA	Frmt Strat	0.94	0.94	0.94	0.91	0.91	0.91	0.91	0.91	0.91
03	REA	Com alpha	0.90	0.89	0.90	0.89	0.90	0.89	0.90	0.89	0.89
04	MAT	Alpha	0.94	0.94	0.94	0.93	0.94	0.93	0.93	0.93	0.94
04	MAT	MC alpha	0.90	0.90	0.90	0.89	0.90	0.89	0.89	0.89	0.90
04	MAT	OR alpha	0.87	0.88	0.88	0.87	0.86	0.86	0.86	0.87	0.88
04	MAT	Frmt Strat	0.94	0.94	0.94	0.94	0.94	0.93	0.94	0.94	0.94
04	MAT	Com alpha	0.93	0.93	0.93	0.93	0.92	0.92	0.92	0.92	0.93
04	REA	Alpha	0.93	0.92	0.92	0.9	0.89	0.89	0.89	0.89	0.89
04	REA	MC alpha	0.92	0.92	0.92	0.89	0.89	0.88	0.88	0.88	0.89
04	REA	OR alpha	0.81	0.80	0.78	0.72	0.72	0.71	0.72	0.71	0.72
04	REA	Frmt Strat	0.93	0.93	0.92	0.90	0.90	0.90	0.90	0.90	0.90
04	REA	Com alpha	0.89	0.89	0.89	0.90	0.89	0.89	0.89	0.89	0.89
05	MAT	Alpha	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
05	MAT	MC alpha	0.87	0.88	0.88	0.88	0.88	0.87	0.88	0.88	0.88
05	MAT	OR alpha	0.89	0.89	0.89	0.89	0.88	0.89	0.88	0.89	0.89
05	MAT	Frmt Strat	0.94	0.94	0.94	0.94	0.94	0.93	0.94	0.94	0.94
05	MAT	Com alpha	0.92	0.92	0.92	0.92	0.92	0.91	0.92	0.92	0.92
05	REA	Alpha	0.92	0.92	0.92	0.88	0.88	0.87	0.87	0.88	0.88
05	REA	MC alpha	0.89	0.89	0.89	0.85	0.84	0.84	0.84	0.85	0.85
05	REA	OR alpha	0.87	0.86	0.86	0.78	0.78	0.78	0.77	0.78	0.79
05	REA	Frmt Strat	0.93	0.93	0.93	0.89	0.89	0.89	0.89	0.89	0.89
05	REA	Com alpha	0.88	0.87	0.88	0.88	0.88	0.87	0.87	0.88	0.88
05	WRI	Alpha	0.72								
05	WRI	MC alpha	0.70								
05	WRI	OR alpha	0.65								
05	WRI	Frmt Strat	0.75								
06	MAT	Alpha	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
06	MAT	MC alpha	0.88	0.89	0.88	0.89	0.88	0.88	0.88	0.89	0.87
06	MAT	OR alpha	0.90	0.90	0.90	0.91	0.90	0.90	0.90	0.90	0.90
06	MAT	Frmt Strat	0.94	0.94	0.94	0.95	0.94	0.94	0.94	0.94	0.94
06	MAT	Com alpha	0.93	0.92	0.93	0.93	0.92	0.92	0.92	0.92	0.92
06	REA	Alpha	0.93	0.92	0.92	0.89	0.89	0.89	0.89	0.89	0.89
06	REA	MC alpha	0.91	0.89	0.89	0.86	0.85	0.85	0.85	0.85	0.85
06	REA	OR alpha	0.88	0.87	0.88	0.83	0.83	0.84	0.83	0.82	0.83
06	REA	Frmt Strat	0.94	0.93	0.93	0.91	0.91	0.91	0.91	0.90	0.91
06	REA	Com alpha	0.89	0.88	0.89	0.89	0.89	0.89	0.89	0.89	0.89

Grade	Subject	Stat	Form01	Form02	Form03	Form04	Form05	Form06	Form07	Form08	Form09
07	MAT	Alpha	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
07	MAT	MC alpha	0.85	0.86	0.86	0.86	0.85	0.86	0.85	0.86	0.86
07	MAT	OR alpha	0.86	0.86	0.87	0.87	0.87	0.87	0.86	0.86	0.87
07	MAT	Frmt Strat	0.92	0.92	0.93	0.93	0.92	0.93	0.92	0.93	0.93
07	MAT	Com alpha	0.91	0.90	0.90	0.91	0.90	0.90	0.90	0.91	0.90
07	REA	Alpha	0.93	0.92	0.93	0.89	0.89	0.89	0.89	0.89	0.89
07	REA	MC alpha	0.90	0.89	0.90	0.85	0.85	0.85	0.85	0.86	0.86
07	REA	OR alpha	0.91	0.90	0.90	0.86	0.86	0.86	0.85	0.86	0.85
07	REA	Frmt Strat	0.94	0.94	0.94	0.91	0.91	0.91	0.91	0.91	0.91
07	REA	Com alpha	0.90	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
08	MAT	Alpha	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.93	0.94
08	MAT	MC alpha	0.89	0.88	0.89	0.89	0.89	0.89	0.88	0.88	0.89
08	MAT	OR alpha	0.90	0.89	0.9	0.89	0.89	0.9	0.9	0.89	0.9
08	MAT	Frmt Strat	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
08	MAT	Com alpha	0.93	0.92	0.92	0.93	0.93	0.93	0.92	0.92	0.93
08	REA	Alpha	0.93	0.93	0.93	0.89	0.89	0.89	0.89	0.88	0.89
08	REA	MC alpha	0.91	0.90	0.90	0.85	0.85	0.85	0.85	0.84	0.84
08	REA	OR alpha	0.92	0.92	0.92	0.88	0.88	0.88	0.88	0.88	0.88
08	REA	Frmt Strat	0.95	0.94	0.94	0.91	0.91	0.91	0.91	0.91	0.91
08	REA	Com alpha	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.88	0.89
08	WRI	Alpha	0.73								
08	WRI	MC alpha	0.64								
08	WRI	OR alpha	0.66								
08	WRI	Frmt Strat	0.75								

Not surprisingly, the reliability is highest when considering the test as a whole, as opposed to a subset of the test. For instance, one would expect the reliability coefficient to be higher when combining the information from all items (all MC/OR) than when considering MC items only or OR items only. This is the pattern that is displayed in Table 7-4.

7.4 Reliability of Achievement Level Categorization

All test scores contain measurement error; thus, classifications based on test scores are also subject to measurement error. After the achievement levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications. For every 2005 NECAP grade and content area, each student is

classified into one of the following achievement levels: Substantially Below Proficient (SBP), Partially Proficient (PP),

Proficient (P), or Proficient With Distinction (PWD). This section of the report explains the methodologies used to assess the reliability of classification decisions, and results are given.

ACCURACY AND CONSISTENCY

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated because errorless test scores do not exist.

Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete, parallel forms of the test are given to the same group of students. In operational assessment programs, however, such a design is usually impractical. To overcome this issue, techniques have been developed to estimate both the accuracy and consistency of classification decisions based on a single administration of a test. The technique developed by Livingston and Lewis (1995) was used for the 2005 NECAP because it is easily adaptable to examinations of all kinds of formats, including mixed-format tests.

CALCULATING ACCURACY

All of the accuracy and consistency estimation techniques used in this section make use of the concept of "true scores" in the sense of classical test theory. A true score is the score that would be obtained on a test that had no measurement error. It is a theoretical concept that cannot be observed, although it can be estimated. In the Livingston and Lewis method, the estimated true scores are used to classify students into their "true" achievement level. After various technical adjustments (which are described in Livingston and Lewis, 1995), a 4×4 contingency table was created for each content area

and grade. The [*i*,*j*] entry of an accuracy table represents the estimated proportion of students whose true score fell into achievement level *i* and whose observed score fell into achievement level *j* on the 2005 NECAP. Overall accuracy, which is the proportion of students whose true and observed achievement levels match one another, is the sum of the diagonal of the accuracy table.

CALCULATING CONSISTENCY

To estimate consistency, the true scores are used to estimate the joint distribution of classifications on two independent, parallel test forms. After statistical adjustments (see Livingston and Lewis, 1995), a new 4 × 4 contingency table was created for each content area and grade that shows the proportion of students who would be classified into each achievement level by the two (hypothetical) parallel test forms. That is, the [*i*,*j*] entry of a consistency table represents the estimated proportion of students whose observed score on the first form would fall into achievement level *i* and whose observed score on the second form would fall into achievement level *j*. Overall consistency, which is the proportion of students classified into exactly the same achievement level by the two forms of the test, is the sum of the diagonal of this new contingency table.

CALCULATING KAPPA

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

$$\kappa = \frac{\text{(Observed agreement) - (Chance agreement)}}{\text{1 - (Chance agreement)}} = \frac{\sum_i C_{ii} - \sum_i C_{i.}C_{.i}}{1 - \sum_i C_{i.}C_{.i}},$$

where:

$C_{i.}$ is the proportion of students whose observed achievement level would be *Level i*, *i*=1,2,3,4, on the first hypothetical parallel form of the test;

$C_{.i}$ is the proportion of students whose observed achievement level would be *Level i*, *i*=1,2,3,4, on the second hypothetical parallel form of the test;

C_{ii} is the proportion of students whose observed achievement level would be *Level i*, *i*=1,2,3,4, on both hypothetical parallel forms of the test.

Because κ is corrected for chance, the values of κ are lower than other consistency estimates.

RESULTS OF ACCURACY, CONSISTENCY, AND KAPPA ANALYSES

Summaries of the accuracy and consistency analyses are provided in Appendix J. This appendix includes the accuracy and consistency contingency tables described above. The overall accuracy and consistency indices are provided as well as the kappa statistic. The overall index is, as described above, the sum of the diagonal elements of the appropriate contingency table.

Accuracy and consistency values conditional upon achievement level are also given in Appendix J. For these calculations, the denominator is the proportion of students who are associated with a given achievement level. For example, the conditional accuracy value is 0.7440 for the PP achievement level for mathematics grade 3. This figure indicates that among the students whose true scores placed them in the PP achievement level, 74.40% of them would be expected to be in the PP achievement level when categorized according to their observed score. Similarly, the corresponding consistency value of 0.6496 indicates that 64.96% of students with observed scores in PP would be expected to score in the PP achievement level again if a second, parallel test form were used.

For certain tests, concern may be greatest regarding decisions made about a particular threshold. For example, if a college gave credit to students who achieved an Advanced Placement test score of four

or five, but not one, two, or three, one might be interested in the accuracy of the dichotomous decision, below four versus four or above. Therefore, Appendix J provides accuracy and consistency results at each of the cut points. These values indicate the overall accuracy and consistency of the dichotomous decisions, either above or below the associated cut point. In addition, the false positive and false negative accuracy rates are shown. These values are estimates of the proportion of students whose observed scores were above the cut despite exhibiting true scores below the cut, and vice versa.

All figures are derived from Livingston & Lewis' (1995) method of estimating the accuracy and consistency of classifications. It should be noted that Livingston & Lewis discuss two versions of the accuracy and consistency tables: a standard version that performs calculations for forms parallel to the form taken, and an "adjusted" version that adjusts the results of one form to match the observed score distribution obtained in the data. In Appendix J, results using the standard version are given. This "unadjusted" version was preferred for two reasons: 1) the unadjusted version can be considered as a smoothing of the data, thereby decreasing the variability of the results; and 2) for results dealing with the consistency of two parallel forms, the unadjusted tables are symmetric, indicating that the two parallel forms have the same statistical properties. The statement in 2) is consistent with the notion of forms that are parallel, i.e., it is more intuitive and interpretable for two parallel forms to have the same statistical distribution as one another.

Descriptive statistics relating to the decision accuracy and consistency of the 2005 NECAP examinations can be derived from Appendix J. For mathematics, overall accuracy ranged from 0.79 to 0.82; overall consistency ranged from 0.71 to 0.75; the kappa statistic ranged from 0.59 to 0.65. For reading, overall accuracy ranged from 0.78 to 0.83; overall consistency ranged from 0.69 to 0.76; the kappa statistic ranged from 0.55 to 0.64. Finally, for writing, overall accuracy was 0.65 or 0.68 in the two grades tested; overall consistency was 0.54 or 0.57; the kappa statistic was 0.36 or 0.38.

Table 7-5 summarizes these results by giving overall, conditional-on-level, and at-cut-point figures for both accuracy and consistency. In Table 7-5, values outside parentheses denote accuracy, while values inside parentheses denote consistency. As in other types of reliability, it is inappropriate to compare results between grades and content areas when analyzing the decision accuracy and consistency of a given examination.

Table 7-5
Summary of Decision Accuracy and Consistency Results

Content/Grade	Overall	Conditional on Level				At Cut Point		
		SBP	PP	P	PWD	SBP:PP	PP:P	P:PWD
Mathematics 3	.82 (.75)	.88 (.81)	.74 (.65)	.84 (.79)	.83 (.73)	.95 (.94)	.93 (.90)	.94 (.92)
Mathematics 4	.81 (.74)	.87 (.79)	.72 (.62)	.83 (.78)	.85 (.75)	.94 (.92)	.92 (.89)	.95 (.93)
Mathematics 5	.80 (.73)	.85 (.77)	.61 (.49)	.85 (.79)	.85 (.76)	.93 (.90)	.92 (.88)	.95 (.93)
Mathematics 6	.81 (.74)	.87 (.80)	.64 (.53)	.85 (.80)	.85 (.76)	.93 (.91)	.92 (.89)	.95 (.94)
Mathematics 7	.79 (.71)	.87 (.80)	.62 (.50)	.82 (.76)	.83 (.72)	.93 (.90)	.91 (.88)	.95 (.92)
Mathematics 8	.81 (.74)	.88 (.82)	.67 (.57)	.84 (.78)	.83 (.72)	.93 (.91)	.92 (.89)	.95 (.93)
Reading 3	.80 (.72)	.83 (.72)	.71 (.61)	.83 (.78)	.83 (.71)	.95 (.93)	.91 (.88)	.94 (.91)
Reading 4	.78 (.70)	.84 (.73)	.71 (.61)	.79 (.73)	.83 (.71)	.94 (.92)	.90 (.86)	.93 (.91)
Reading 5	.78 (.69)	.82 (.70)	.69 (.59)	.81 (.75)	.82 (.69)	.94 (.92)	.90 (.86)	.94 (.92)
Reading 6	.81 (.73)	.84 (.73)	.73 (.64)	.84 (.79)	.83 (.70)	.95 (.93)	.91 (.87)	.95 (.93)
Reading 7	.81 (.73)	.84 (.74)	.74 (.64)	.84 (.79)	.82 (.69)	.95 (.93)	.91 (.87)	.95 (.93)
Reading 8	.83 (.76)	.86 (.77)	.77 (.69)	.85 (.81)	.84 (.73)	.95 (.93)	.92 (.89)	.96 (.94)
Writing 5	.65 (.54)	.77 (.61)	.57 (.48)	.67 (.58)	.73 (.50)	.89 (.84)	.83 (.77)	.93 (.89)
Writing 8	.68 (.57)	.77 (.61)	.58 (.49)	.73 (.65)	.71 (.42)	.89 (.84)	.84 (.78)	.95 (.93)

CHAPTER 8—VALIDITY

The purpose of this report is to describe several technical aspects of the 2005 NECAP tests in an effort to assess the validity evidence to support NECAP score interpretations. Because it is the interpretations of test scores that are evaluated for validity, not the test itself, this report presents documentation to evaluate intended interpretations (AERA, 1999). Each of the chapters in this report contributes important information to the investigation of validity by addressing the following aspects of NECAP: test development and design; test administration; scoring, scaling, and equating; item analyses; reliability; and score reporting.

The NECAP assessments are based on and aligned with the content standards and performance indicators in the GLEs for mathematics, reading, and writing. Intended inferences from the NECAP results are about student achievement on these content standards, and such achievement inferences are meant to be useful for program and instructional improvement and as a component of school accountability.

The *Standards for Educational and Psychological Testing* (1999) provides a framework for describing sources of evidence that should be considered when constructing an assessment of validity. These sources include evidence based on the following five general areas: test content, response processes, internal structure, consequences of testing, and relationship to other variables. Although each of these sources may speak to a different *aspect* of validity, they are not distinct *types* of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

A measure of test content validity is to determine how well the assessment tasks represent the curriculum and standards for each subject and grade level. This is informed by the item development

process, including how the test blueprints and test items align with the curriculum and standards. Viewed through this lens provided by the Standards, evidence based on test content was extensively described in Chapter 2. Item alignment with content standards; item bias; sensitivity and content appropriateness review processes; adherence to the test blueprint; use of multiple item types; use of standardized administration procedures, with accommodated options for participation; and appropriate test administration training are all components of validity evidence based on test content. As discussed earlier, all NECAP test questions are aligned by educators with specific content standards and undergo several rounds of review for content fidelity and appropriateness. Items are presented to students in multiple formats (MC, SA, and CR). Finally, tests are administered according to mandated standardized procedures, with allowable accommodations, and all test coordinators and test administrators are required to familiarize themselves with and adhere to all of the procedures outlined in the *NECAP Test Coordinator* and *Test Administrator* manuals.

The scoring information in Chapter 4 describes both the steps taken to train and monitor hand-scorers and quality control procedures related to scanning and machine-scoring. To speak to student response processes, however, additional studies may be helpful and might include an investigation of students' cognitive methods using think-aloud protocols.

Evidence based on internal structure is presented in great detail in the discussions of scaling and equating, item analyses, and reliability in Chapters 5, 6, and 7. Technical characteristics of the internal structure of the assessments are presented in terms of classical item statistics (item difficulty, item-test correlation), differential item functioning analyses, and a variety of reliability coefficients, SEM, and IRT parameters and procedures. In general, item difficulty and discrimination indices were in acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that students who performed well on

individual items tended to perform well overall. Chapter 5 also notes that multiple equating methods will be considered for the equating of future administrations to the 2005 scale.

Evidence based on the consequences of testing is addressed in the scaled scores and reporting information in Chapters 5 and 9 and in the *Guide to Using the 2005 NECAP Reports*, which is a separate document that is referenced in the discussion of reporting. Each of these chapters speaks to the efforts undertaken to promote accurate and clear information to the public regarding test scores. Scaled scores offer the advantage of simplifying the reporting of results across content areas, grade levels, and subsequent years. Achievement levels provide users with reference points for mastery at each grade level, which is another useful and simple way to interpret scores. Several different standard reports are provided to stakeholders. Additional evidence of the consequences of testing could be supplemented with broader investigation of the impact of testing on student learning.

8.1 SUMMARY OF VALIDITY EVIDENCE

The text in the beginning of this chapter offers a detailed description of the measures taken to ensure the validity of the 2005 NECAP score interpretations. Subsections 8.1 and 8.2 supplement this information by presenting an empirical assessment of validity based on the 2005 NECAP data. The empirical assessment provides evidence of *external validity*, which concerns the concordance between test scores and another numerical variable that is expected to be correlated with test scores. For NECAP, external validity is conveyed by the relationship of test scores and situational variables, such as teacher judgments and questionnaire data. The former of these situational variables is examined in this subsection; the latter is considered in subsection 8.2.

Teacher judgments were made about student proficiency in all three content areas: mathematics, reading, and writing. All judgments were made by the student's teachers in the appropriate content area

and were collected on student answer booklets after testing was completed. In each content area, the teacher judgment could fall into 12 possible categories, with three categories corresponding to each NECAP achievement level. The convention of using three categories for each achievement level provided a means for teachers to distinguish among students within the same achievement level. For purposes of analysis, however, the categories were collapsed back into the four achievement levels defined for NECAP: *Substantially Below Proficient (SBP)*, *Partially Proficient (PP)*, *Proficient (P)*, and *Proficient With Distinction (PWD)*. Collapsing teacher judgments into the four NECAP achievement levels allowed for a direct evaluation of the concordance between teacher judgments and test classifications.

The assessment of external validity focused on the relation between the teacher judgment of a student and the student's observed 2005 NECAP achievement level in the same content area. A positive correlation between teacher judgments and NECAP classifications would provide evidence for the external validity of NECAP score interpretations. However, it was not expected that the two variables would align exactly. As discussed in Chapter 7, any measurement includes some amount of measurement error; this is true of both academic assessments and teacher judgments, the latter of which are human ratings and thus have implicit variability of their own. The teacher judgments were thus not viewed as "true" student achievement levels, with which the observed NECAP classifications should match as closely as possible; rather, investigation of the concordance between the two variables was merely designed as a reasonability check of external validity. Nevertheless, it was expected that the majority of teacher judgments would fall within one achievement level of the corresponding NECAP classification.

Tables K-1 to K-14 in Appendix K give cross-tabulations of teacher judgments and observed NECAP achievement levels. The [*i, j*] entry of a table shows the number and percentage of students

whose teacher judgment was achievement level *i* and whose observed NECAP classification was achievement level *j*. For each cell of a given table, the number outside parentheses represents the count of students who fell into that cell; the number inside parentheses gives the corresponding percentage. The *exact or adjacent agreement rates* were also calculated for each grade and content area; these denote the percentage of students whose teacher judgments were either identical or adjacent to their observed NECAP achievement levels. For mathematics, exact or adjacent agreement rates ranged from 92.86% to 94.47%; for reading, they ranged from 95.56% to 96.50%; for writing, they were 93.64% and 94.08% in the two grades tested.

8.2 QUESTIONNAIRE DATA

Another measure of external validity was provided by comparing student scores and achievement levels with answers to a questionnaire that was administered at the end of examination. The questionnaire contained 26 questions: Eight concerned the content area of reading, nine concerned mathematics, and nine concerned writing. Most of the questions were designed for the purpose of gathering information about students and their study habits; however, a subset could be utilized in the assessment of external validity. One question was chosen from each content area that was most expected to correlate with student performance on NECAP examinations. To the extent that the answers to those questions did correlate with student performance in the anticipated manner, the external validity of score interpretations was confirmed. The three identified questions were Questions 7 (reading), 13 (mathematics), and 26 (writing). They will be discussed one at a time.

Question 7 is shown below:

7. How often do you choose to read in your free time?
   A. almost every day
   B. a few times a week
   C. a few times a month
   D. I almost never read.

It was anticipated that students who read more in their free time would have higher average scaled scores and achievement level designations in reading than students who did not read as much. In particular, it was expected that on average, reading performance among students who chose "A" would meet or exceed performance of students who chose "B," whose performance would meet or exceed that of students who chose "C," whose performance would meet or exceed that of students who chose "D." This pattern is displayed in Table 8-1 for all grades, both in terms of average scaled scores and the percentage of students in the *Proficient with Distinction* achievement level.

Table 8-1
Item 7 of Student Questionnaire
Reading

Grade	Response	NResp	%Resp	AvgSS	NSBP	NPP	NP	NPWD	%SBP	%PP	%P	% PWD
03		2984	9	341	620	655	1325	384	21	22	44	13
	A	16197	50	347	1634	3037	8332	3194	10	19	51	20
	B	8123	25	345	932	1618	4317	1256	11	20	53	15
	C	2011	6	342	363	469	973	206	18	23	48	10
	D	2813	9	340	593	719	1316	185	21	26	47	7
04		2675	8	440	551	606	1227	291	21	23	46	11
	A	15699	48	445	1594	2854	8582	2669	10	18	55	17
	B	9592	29	443	1125	2121	5262	1084	12	22	55	11
	C	2062	6	441	380	466	1062	154	18	23	52	7
	D	2534	8	438	596	752	1100	86	24	30	43	3
05		2491	7	539	540	681	1025	245	22	27	41	10
	A	15163	45	546	1379	2882	7921	2981	9	19	52	20
	B	10591	32	543	1234	2696	5462	1199	12	25	52	11
	C	2570	8	541	415	705	1239	211	16	27	48	8
	D	2616	8	537	635	890	1009	82	24	34	39	3
06		3092	9	639	717	830	1291	254	23	27	42	8
	A	12354	36	648	922	2202	6795	2435	7	18	55	20
	B	11932	34	644	1293	3233	6250	1156	11	27	52	10
	C	3850	11	641	563	1131	1921	235	15	29	50	6
	D	3494	10	638	744	1217	1443	90	21	35	41	3
07		3942	11	739	927	1091	1660	264	24	28	42	7
	A	10000	28	749	727	1576	5803	1894	7	16	58	19
	B	11358	32	744	1228	2767	6285	1078	11	24	55	9
	C	4636	13	742	628	1381	2362	265	14	30	51	6
	D	5314	15	738	1088	1847	2236	143	20	35	42	3
08		3345	9	838	870	952	1256	267	26	28	38	8
	A	9843	27	849	650	1648	5527	2018	7	17	56	21
	B	10930	30	844	1134	2815	5821	1160	10	26	53	11
	C	5809	16	841	774	1855	2835	345	13	32	49	6
	D	6661	18	838	1353	2424	2687	197	20	36	40	3

Question 13, which was selected for the assessment of external validity for mathematics, is displayed below:

13. How often do you have mathematics homework?
    A. almost every day
    B. a few times a week
    C. a few times a month
    D. I usually don't have homework in mathematics.

As anticipated, the concurrence between Question 13 and student performance in mathematics (Table 8-2) mirrored the pattern of Question 7 at each grade: On average, mathematics performance among students who chose "A" met or exceeded the performance of students who chose "B," whose performance met or exceeded that of students who chose "C," whose performance met or exceeded that

of students who chose "D." This pattern was again evident both in terms of average scaled scores and the percentage of students in the *Proficient With Distinction* achievement level.

Table 8-2
Item 13 of Student Questionnaire
Mathematics

Grade	Response	NResp	%Resp	AvgSS	NSBP	NPP	NP	NPWD	%SBP	%PP	%P	% PWD
03		3021	9	339	835	674	1142	370	28	22	38	12
	A	13810	43	343	2250	2900	6102	2558	16	21	44	19
	B	10936	34	343	1494	2423	5166	1853	14	22	47	17
	C	2210	7	343	360	456	1033	361	16	21	47	16
	D	2284	7	340	510	612	939	223	22	27	41	10
04		2611	8	438	719	600	969	323	28	23	37	12
	A	15628	48	443	2461	3244	7319	2604	16	21	47	17
	B	10693	33	443	1742	2472	4939	1540	16	23	46	14
	C	2117	6	442	393	452	971	301	19	21	46	14
	D	1659	5	438	469	432	637	121	28	26	38	7
05		2508	7	538	813	528	914	253	32	21	36	10
	A	17769	53	543	3118	3336	8285	3030	18	19	47	17
	B	10283	31	542	2099	2117	4621	1446	20	21	45	14
	C	1768	5	542	394	356	776	242	22	20	44	14
	D	1235	4	537	446	265	426	98	36	21	34	8
06		3017	9	637	986	624	1070	337	33	21	35	11
	A	19029	55	643	3342	3814	8767	3106	18	20	46	16
	B	10594	30	641	2375	2283	4636	1300	22	22	44	12
	C	1199	3	639	356	253	447	143	30	21	37	12
	D	1027	3	634	471	184	293	79	46	18	29	8
07		3845	11	736	1397	800	1284	364	36	21	33	9
	A	20698	59	742	3801	4160	9418	3319	18	20	46	16
	B	8903	25	740	2320	2004	3603	976	26	23	40	11
	C	944	3	735	395	191	304	54	42	20	32	6
	D	951	3	732	483	185	234	49	51	19	25	5
08		3379	9	835	1341	650	1059	329	40	19	31	10
	A	22545	62	842	4097	4680	10113	3655	18	21	45	16
	B	8579	23	838	2591	2158	3187	643	30	25	37	7
	C	1003	3	834	437	221	295	50	44	22	29	5
	D	1136	3	830	645	211	239	41	57	19	21	4

Finally, Question 26, which was selected for the investigation of writing, is shown below:

26. What kinds of writing do you do most in school?
    A. I mostly write stories.
    B. I mostly write reports.
    C. I mostly write about things I've read.
    D. I do all kinds of writing.

Unlike Questions 7 and 13, no distinction was made between choices "A," "B," and "C" in terms of expected writing performance on NECAP. The only anticipated outcome was that students who selected

choice "D," ostensibly having had experience in many different kinds of writing, would tend to outperform students who selected any other answer choice. For both grades 5 and 8, this outcome was realized both in terms of average scaled score and the percentage of students in the *Proficient with Distinction* achievement level (Table 8-3).

Table 8-3
Item 26 of Student Questionnaire
Writing

Grade	Response	NResp	%Resp	AvgSS	NSBP	NPP	NP	NPWD	%SBP	%PP	%P	% PWD
05		3510	11	535	822	1169	1207	312	23	33	34	9
	A	6157	18	538	996	2206	2468	487	16	36	40	8
	B	3380	10	537	619	1301	1231	229	18	38	36	7
	C	2904	9	538	527	975	1172	230	18	34	40	8
	D	17456	52	541	2039	5212	7824	2381	12	30	40	14
08		4485	12	834	1235	1470	1534	246	28	33	34	5
	A	3941	11	835	957	1539	1345	100	24	39	34	3
	B	5611	15	837	1070	2034	2279	228	19	36	41	4
	C	4149	11	837	830	1540	1576	203	20	37	38	5
	D	18265	50	841	2020	5637	9095	1513	11	31	50	8

Based on the foregoing analysis, the relation between questionnaire data and performance on the NECAP was consistent with expectations for all three questions selected for the investigation of external validity. See Appendix L for a copy of the questionnaire in its entirety, and see Tables L-1 to L-14 in Appendix L for complete data on the concordance between questionnaire items and test performance. For all grades and content areas, Tables L-1 to L-14 give the following information about each possible questionnaire item and each possible answer choice:

- N count and percentage of students who selected the answer choice
- Number and percentage of students in each achievement level, among students who selected the answer choice
- Average scaled score among students who selected the answer choice.

8.3 VALIDITY STUDIES AGENDA

An assessment of the statistical evidence of validity of 2005 NECAP score interpretations is presented earlier in this chapter, subsections 8.1 and 8.2. The remaining part of this chapter describes further studies of validity that are being considered for the future. These studies could enhance the investigations of validity that have already been performed. The proposed areas of validity to be examined fall into four categories: *external validity, convergent and discriminant validity, structural validity, and procedural validity*. These will be discussed in turn.

EXTERNAL VALIDITY

For the 2005 NECAP score interpretations, external validity was assessed through cross-tabulations of NECAP test scores with teacher judgments and questionnaire data. Future investigations could involve additional variables with which to correlate NECAP results. For example, data could be collected on the grades of each student who took the NECAP examinations. As with the analysis of teacher judgments and questionnaire data, cross-tabulations of NECAP achievement levels and assigned grades could be created. The average NECAP scaled score could also be computed for each possible assigned grade (A, B, C, etc.). Analysis would focus on the concordance between NECAP scores and grades in the appropriate class (e.g., NECAP mathematics would be correlated with student grades in mathematics, not reading). NECAP scores could also be correlated with other appropriate classroom assessments in addition to final grades.

Another potential examination of external validity would be to correlate NECAP scores with scores on another standardized test, such as the Iowa Test of Basic Skills (ITBS). As with the study of concordance between NECAP scores and grades, this investigation would compare scores in analogous content areas (e.g., NECAP reading and ITBS reading comprehension). All tests taken by each student would be appropriate to the student's grade level. Initial analyses of the school-level correlations

between the 2005 NECAP tests and previous state assessments administered through 2004 show a strong relationship between the tests for schools with 45 or more students. At the elementary level, correlations range form .70 to .94. Correlations at the middle school level range from .85 to .94.

CONVERGENT AND DISCRIMINANT VALIDITY

The concepts of convergent and discriminant validity were defined by Campbell and Fiske (1959) as specific types of validity that fall under the umbrella of *construct validity*. The notion of convergent validity states that measures or variables that are intended to align with one another should actually be aligned in practice. Discriminant validity, on the other hand, is the idea that measures or variables that are intended to differ from one another should not be too highly correlated. Assessment of validity is accomplished by comparing the correlations of variables that are intended to correlate with those that are not.

Campbell and Fiske (1959) introduced the study of different *traits* and *methods* as the means of assessing convergent and discriminant validity. Traits refer to the constructs that are being measured (e.g., mathematical ability), and methods are the instruments of measuring them (e.g., a mathematics examination or grade). To utilize the framework of Campbell and Fiske, it is necessary that more than one trait and more than one method be examined. Analysis is performed through the multitrait-multimethod matrix, which gives all possible correlations of the different combinations of traits and methods. Campbell and Fiske define four properties of the multitrait-multimethod matrix that serve as evidence of convergent and discriminant validity:

- The correlation among different methods of measuring the same trait should be sufficiently different from zero. For example, scores on a mathematics examination and grades in a mathematics class should be positively correlated.

- The correlation among different methods of measuring the same trait should be higher than different methods of measuring different traits. For example, scores on a mathematics examination and grades in a mathematics class should be more highly correlated than scores on a mathematics examination and grades in a reading class.
- The correlation among different methods of measuring the same trait should be higher than the same method of measuring different traits. For example, scores on a mathematics examination and grades in a mathematics class should be more highly correlated than scores on a mathematics examination and scores on an analogous reading examination.
- The pattern of correlations should be similar across comparisons of different traits and methods. For example, if the correlation between test scores in reading and writing is higher than the correlation between test scores in reading and mathematics, it is expected that the correlation between grades in reading and writing would also be higher than the correlation between grades in reading and mathematics.

For NECAP, convergent and discriminant validity could be examined by constructing a multitrait-multimethod matrix and analyzing the four pieces of evidence described above. The traits examined would be mathematics, reading, and writing; different methods would include NECAP score and such variables as grades, teacher judgments, and/or scores on another standardized test.

STRUCTURAL VALIDITY

Though the previous types of validity examine the concurrence between different measures of the same content area, structural validity focuses on the relation between strands *within* a content area, thus supporting *content validity*. Standardized tests are carefully designed to ensure that all appropriate strands of a content area are adequately covered in examination, and structural validity is the degree to which related elements of a test are correlated in their intended manner. For instance, it is desired that

performance on different strands of a content area be positively correlated; however, as these strands are designed to measure distinct components of the content area, it is reasonable to expect that each strand would contribute a unique component to the assessment. Additionally, it is desired that the correlation between different item types (MC, SA, and CR) of the same content area be positive.

As an example, an analysis of NECAP structural validity would investigate the correlation between performance in Geometry and Measurement and performance in Functions and Algebra. Additionally, the concordance between performance on MC items and OR items would be examined. Such a study would address the consistency of NECAP examinations within each grade and content area.

PROCEDURAL VALIDITY

As mentioned earlier, the *NECAP Test Coordinator* and *Test Administrator* manuals delineate the procedures to which all NECAP test coordinators and test administrators are required to adhere. A study of procedural validity would provide a comprehensive documentation of the procedures that were followed throughout the NECAP administration. The results of the documentation would then be compared to the manuals, and procedural validity would be confirmed to the extent that the two are in alignment. Evidence of procedural validity is important because it verifies that the actual administration practices are in accord with the intentions of the design.

Possible instances where discrepancies can exist between design and implementation include the following: A teacher may spiral test forms incorrectly within a classroom; cheating may occur among students; or answer documents may be scanned incorrectly. These are examples of *administration error*. A study of procedural validity involves capturing any administration errors and presenting them within a cohesive document for review.

All potential examinations of validity that have been introduced in this chapter will be discussed as candidates for action by the NECAP Technical Advisory Committee (NECAP TAC) during 2006-2007. With the advice of the NECAP TAC, the states will develop a short-term (e.g., 1-year) and longer term (e.g., 2-year to 5-year) plan for validity studies.

SECTION III: NECAP REPORTING

CHAPTER 9—SCORE REPORTING

9.1 TEACHING YEAR VS. TESTING YEAR REPORTING

The data used for the NECAP Reports are the results of the fall 2005 administration of the NECAP test. However, the NECAP tests are based on the GLEs from the prior year. For example, the Grade 7 NECAP test, administered in the fall of seventh grade, is based on the grade 6 GLEs. Therefore, many students receive the instruction they need for this fall test at a school different from where they are currently enrolled. The state Departments of Education determined that it would be valuable for both the school where the student tested and the school where the student received instruction to have access to information that can help improve to curriculum. To achieve this goal, separate Item Analysis, School and District Results, and School and District Summary reports were created for the "testing" school and the "teaching" school. Every student who participated in the NECAP test will be represented in "testing" reports, and most students will also be represented in "teaching" reports. In some instances, such as when the student has recently moved into the state, it is not possible to provide information about a student in "teaching" reports.

9.2 PRIMARY REPORTS

There were four primary reports for the 2005–06 NECAP:

- Student Report
- Item Analysis Report
- School and District Results Report
- School and District Summary Report

With the exception of the Student Reports, all reports were available for schools and districts to view or download on a password secure Web site hosted by Measured Progress. Student-level data files were also available for districts to download from the secure Web site. Each of these reports is described in the following subsections. Sample reports are provided in Appendix M.

9.3 STUDENT REPORT

The *NECAP Student Report* is a single-page two-sided report that is divided into three sections. The front side of the report includes a letter from the commissioner of education, a description of the achievement levels, and a graph showing state summary results. The reverse side of the student report provides a complete picture of an individual student's performance on the NECAP and is divided into three sections. The first section gives the student's overall performance for each content area. The student's achievement levels and scaled scores are shown, both in a table and graphically. The graphic display shows the range of possible scaled scores divided into the four achievement levels. There is also a display of the standard error of measurement bar for each content area.

The second section of the report displays the student's achievement level relative to the percent of students in each achievement level for the school, district, and state for each content area.

The third section of the report shows the student's performance compared to school, district, and statewide performance in a variety of areas. Each of the three content areas assessed by NECAP is reported by subcategories. For **reading**, with the exception of Word ID/Vocabulary items, items are reported in two ways: Type of Text and Level of Comprehension. The two types of text are Literary and Informational. The two levels of comprehension are Initial Understanding and Analysis and Interpretation. Numbers and Operations, Geometry and Measurement, Functions and Algebra, and Data, Statistics, and Probability are the subcategories reported for **mathematics**. The content area subcategories for **writing** are reported on the Structures of Language and Writing Conventions, displayed in the student's writing and in response to MC items, and by the type of response—short or extended.

Student performance in all content area subcategories is presented as a table including possible points; points earned by this student; average points earned for the school, district, and state; and the average points earned by students at the Proficient level on the total content area test.

To provide a more complete picture of this student's performance on the writing assessment in grades 5 and 8, each scorer chose up to three comments about the student's writing performance from a predetermined list produced by the writing representatives from each state department of education. The comments selected by the student's scorers appear in the table at the bottom right hand corner of the report.

The *NECAP Student Report* is confidential and should be kept secure within the school and district. The Family Educational Rights and Privacy Act (FERPA) requires that access to individual student results be restricted to the student, the student's parents/guardians, and authorized school personnel.

9.4 ITEM ANALYSIS REPORTS

The *NECAP Item Analysis Report* provides a roster of all the students in each school and shows their performance on the common items in the assessment. One report is provided for each content area. The student names are listed down the left side of the report, and the items are listed across the top in the order in which they appear in the released item documents (not the position in which they appeared on the test). For each item, the following seven pieces of information are provided: the released item number, the content strand for the item, the GLE code for the item, the Depth of Knowledge code for the item, the item type, the correct response letter for MC items, and the total possible points for each item. For each student, each MC item is marked either with a plus sign (+), indicating that the student chose the correct response, or a letter (from A to D), indicating which incorrect response the student chose. For CR items, the number of points that the student attained is shown. All responses to released items are shown is the report, regardless of the student's participation status.

The columns on the right side of the report show Total Test Results, which are broken into several categories. The Subcategory Points Earned columns report the total possible points and total points that the student earned in each content strand. The Total Points Earned column is a summary of all of the total possible points and points earned by the student in each of the content areas. The last two columns show the Scaled Score and Achievement Level for each student. For students who are reported as Not Tested, a code appears in the Achievement Level column to indicate the reason why the student did not test. The descriptions of these codes can be found on the legend, after the last page of data on the report. It is important to note that not all items used to compute student scores are included in this report. Only those items that have been released are included. At the bottom of the report, the average percent correct for each MC item and average score for the SA and CR items and writing prompts is shown for the school, district, and state.

The *NECAP Item Analysis Report* is confidential and should be kept secure within the school and district. The FERPA requires that access to individual student results be restricted to the student, the student's parents/guardians, and authorized school personnel.

9.5 SCHOOL AND DISTRICT RESULTS REPORTS

The *NECAP School Results Report* and the *NECAP District Results Report* consist of three parts: the Grade Level Summary Report (page 2), the Content Area Results (pages 3, 5, and 7), and the Disaggregated Content Area Results (pages 4, 6, and 8).

The Grade Level Summary Report provides a summary of participation in the NECAP and a summary of NECAP results. The Summary of Participation section on the top half of the page shows the number and percentage of students who were enrolled as of October 1, 2005. The total number of students reported as enrolled is defined as the number of students tested added to the number of students that were not tested.

Because students who were not tested did not participate in the NECAP tests, average school scores are not affected by non-tested students. These students are included in the calculation of the percent that participated but are not included in the calculation of scores. For students who participated in some but not all sessions of the NECAP test, their actual score was reported for each content area in which they participated. These reporting decisions were made to support the requirement that all students must participate in the NECAP testing program.

Data are provided for the following groups of students who may not have completed the entire battery of NECAP tests:

- **Alternate Assessment**: Students in this category completed an alternate assessment for the 2004–2005 school year.

- **First-Year LEP**: Students in this category are defined as being new to the United States after October 1, 2004 and were not required to take the NECAP tests in reading and writing. Students in this category were expected to take the mathematics portion of the NECAP.
- **Withdrew After October 1**: Students withdrawing from a school after October 1, 2005 may have taken some sessions of the NECAP tests prior to their withdrawal from the school.
- **Enrolled After October 1**: Students enrolling in a school after October 1, 2005 may not have had adequate time to fully participate in all sessions of the NECAP tests.
- **Special Consideration**: Schools received state approval for special consideration for an exemption for all or part of the NECAP tests for any student whose circumstances are not described by the previous categories but for whom the school determined that taking the NECAP tests would not be possible.
- **Other**: Occasionally students will not have completed the NECAP tests for reasons other than those listed above. These "other" categories were considered "not state approved."

The Summary of Results section on the bottom half of the page shows the number and percentage of students performing at each achievement level in each of the three content areas for the school, district, and state. In addition, a Mean Scaled Score is provided for each content area at the school, district, and state levels. For the district version of this report, the school information is blank.

The Content Area Results pages provide information on performance in specific subcategories of the tested content areas (for example, geometry, and measurement within mathematics). Subscore results by content area tested are provided on the following pages of the report:

- page 3: reading
- page 5: mathematics
- page 7: writing

The purpose of these sections is to help schools to determine the extent to which their curricula are effective in helping students to achieve the particular standards and benchmarks contained in the *Grade Level Expectations*. Information about each content area (reading, mathematics and writing) for school, district, and state includes

- the total number of students Enrolled, Not Tested (state-approved reason), Not Tested (other reason), and Tested;
- the total number and percent of students at each achievement level (based on the number in the Tested column); and
- the Mean Scaled Score.

This information is provided for 2005–06 testing year. These pages of the report have been designed both to include a location for 2006–07 and 2007–08 scores and so that in the future the cumulative average over the three years can be reported. Again, for the district version of this report, the school information is blank.

Information about each content area subcategory for reading, mathematics and writing includes the following:

- The **Total Possible Points** for that category. In order to provide as much information as possible for each category, the total number of points includes both the common items used to calculate scores and additional items in each category used for equating the test from year to year.
- A graphic display of the **Percent of Total Possible Points** for the school, state, and district. In this graphic display, there are symbols representing school, district, and state performance. In addition, there is a line representing the standard error of measurement. This statistic

indicates how much a student's score could vary if the student were examined repeatedly with the same test (assuming that no learning occurs between test administrations).

The Disaggregated Content Area Results pages present the relationship between the variables reported and performance in each content area at the school, district, and state levels. Disaggregated Results for each content area are shown on the following pages of the report:

- page 4: reading
- page 6: mathematics
- page 8: writing

Each content area page shows the number of students categorized as Enrolled, Not Tested (state-approved reason), Not Tested (other reason), and Tested. The tables also provide the number and percentage of students within each of the four achievement levels and the Mean Scaled Score by each reporting category.

Below is a list of the reporting categories:

- gender
- Primary Race/Ethnicity
- Limited English Proficiency (LEP)
- IEP
- socioeconomic status (SES)
- migrant
- Title I
- 504 Plan

The data for achievement levels and Mean Scaled Score are based on the number shown in the Tested column. The data for the reporting categories were provided by information coded on the

students' answer booklets by teachers and/or data linked to the student label. Because performance is being reported by categories that can contain relatively low numbers of students, school personnel are advised, under FERPA guidelines, to treat these pages confidentially.

Please note that for NH and VT, no data were reported for the 504 Plan in any of the content areas. In addition, for VT, no data were reported for Title I in any of the content areas.

9.6 SCHOOL AND DISTRICT SUMMARY REPORTS

The *NECAP School Summary Report* and the *NECAP District Summary Report* provide details, broken down by content area, about student performance for all grade levels of NECAP that were tested in the school. The purpose of the summary is to help schools to determine the extent to which their students achieve the particular standards and benchmarks contained in the *Grade Level Expectations.* Information about each content area and grade level for school, district, and state includes

- the total number of students Enrolled, Not Tested (state-approved reason), Not Tested (other reason), and Tested
- the total number and percent of students at each achievement level (based on the number in the Tested column) and
- the Mean Scaled Score.

The data reported, report format, and guidelines for using the reported data are identical for both the school and district reports. The only real difference between the reports is that the *NECAP District Summary Report* includes no individual school data. A separate school report and district report was produced for each grade level tested.

9.7 DECISION RULES

To ensure that reported results for the 2005–06 NECAP are accurate relative to collected data and other pertinent information, a document that delineates analysis and reporting rules was created. These decision rules were observed in the analyses of NECAP test data and in reporting the assessment results. Moreover, these rules are the main reference for quality assurance checks.

The decision rules document used for reporting results of the October 2005 administration of the NECAP is founded in Appendix N.

The first set of rules pertains to general issues in reporting scores. Each issue is described, and pertinent variables are identified. The actual rules applied are described by the way they impact analyses and aggregations and their specific impact on each of the reports. The general rules are further grouped into issues pertaining to test items, school type, student exclusions, and number of students for aggregations.

The second set of rules pertains to reporting student participation. These rules describe which students were counted and reported for each subgroup in the student participation report.

9.8 QUALITY ASSURANCE

Quality assurance measures are embedded throughout the entire process of analysis and reporting. The data processor, data analyst, and psychometrician assigned to work on the NECAP implement quality control checks of their respective computer programs and intermediate products. Moreover, when data are handed off to different functions within the Research and Analysis division, the sending function verifies that the data are accurate before handoff. Additionally, when a function receives a data set, the first step is to verify the data for accuracy.

Another type of quality assurance measure is parallel processing. Students' scaled scores for each content area are assigned by a psychometrician through a process of equating and scaling. The scaled scores are also computed by a data analyst to verify that scaled scores and corresponding achievement levels are assigned accurately. Respective scaled scores and achievement levels assigned are compared across all students for 100% agreement. Different exclusions assigned to students that determine whether each student receives scaled scores and/or is included in different levels of aggregation are also parallel-processed. Using the decision rules document, two data analysts independently write a computer program that assigns students' exclusions. For each subject and grade combination, the exclusions assigned by each data analyst are compared across all students. Only when 100% agreement is achieved can the rest of data analysis be completed.

The third aspect of quality control involves the procedures implemented by the quality assurance group to check the veracity and accuracy of reported data. Using a sample of schools and districts, the quality assurance group verifies that reported information is correct. The step is conducted in two parts: (1) verify that the computed information was obtained correctly through appropriate application of different decision rules and (2) verify that the correct data points populate each cell in the NECAP reports. The selection of sample schools and districts for this purpose is very specific and can affect the success of the quality control efforts. There are two sets of samples selected that may not be mutually exclusive. The first set includes those that satisfy the following criteria:

- One-school district
- Two-school district
- Multi-school district

The second set of samples includes districts or schools that have unique reporting situations as indicated by decision rules. This set is necessary to check that each rule is applied correctly. The second set includes the following criteria:

- Private school
- Small school that receives no school report
- Small district that receives no district report
- District that receives a report but all schools are too small to receive a school report
- School with excluded (not tested) students
- School with home-schooled students

The quality assurance group uses a checklist to implement its procedures. After the checklist is completed, sample reports are circulated for psychometric checks and program management review. The appropriate sample reports are then presented to the client for review and sign-off.

SECTION IV: REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). Fort Worth: Holt, Rinehart and Winston.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23,* 355–368.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). New York: John Wiley & Sons, Inc.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105–146). New York: Macmillan Publishing Co.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Hambleton, R. K., & van der Linden, W. J. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Joint Committee on Testing Practices (1988). *Code of Fair Testing Practices in Education*. Washington, D.C.: National Council on Measurement in Education.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32,* 179–197.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.